Titre: Title:	Amélioration de la gestion des réservations de surplus en utilisant l'intelligence artificielle
Auteur: Author:	Bessem Dammak
Date:	2025
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Dammak, B. (2025). Amélioration de la gestion des réservations de surplus en utilisant l'intelligence artificielle [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/66603/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/66603/
Directeurs de recherche: Advisors:	Soumaya Yacout, & Antoine Saucier
Programme: Program:	Maîtrise recherche en génie industriel

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Amélioration de la gestion des réservations de surplus en utilisant l'intelligence artificielle

BESSEM DAMMAK
Département de mathématiques et de génie industriel

Mémoire présentée en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie industriel

Juin 2025

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Amélioration de la gestion des réservations de surplus en utilisant l'intelligence artificielle

présenté par **Bessem DAMMAK**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Julie CARREAU, présidente
Soumaya YACOUT, membre et directrice de recherche
Antoine SAUCIER, membre et codirecteur de recherche
Safa BHAR LAYEB, membre

DÉDICACE

À mes parents, pour leur soutien et leurs encouragements constants
À mes sœurs, pour leur appui tout au long de ce parcours
À mes chers amis,

REMERCIEMENTS

Je souhaite tout d'abord exprimer ma profonde gratitude à Madame Soumaya YACOUT, ma directrice de recherche, pour son accompagnement précieux tout au long de cette maîtrise. Son engagement, sa rigueur scientifique et la qualité de ses retours m'ont permis de progresser tant sur le plan méthodologique que personnel. Ses conseils avisés ont été déterminants dans l'avancement de ce projet et dans l'atteinte des résultats obtenus.

Je tiens également à remercier chaleureusement Monsieur Antoine SAUCIER, codirecteur de recherche, pour sa disponibilité constante et ses orientations pertinentes. Ses commentaires constructifs et sa vision critique ont enrichi cette recherche et m'ont guidé dans les différentes étapes de ce parcours académique.

Je tiens également à adresser mes sincères remerciements à Madame Julie CARREAU, présidente du jury, ainsi qu'à Madame Safa BHAR LAYEB, membre du jury, pour leur précieuse participation à l'évaluation de ce mémoire. Je suis particulièrement honoré de bénéficier de leur regard éclairé et de leur grande expertise.

RÉSUMÉ

Les programmes de fidélisation jouent un rôle central dans la gestion des marques, en renforçant leur notoriété et en favorisant des relations durables avec la clientèle. Ces programmes offrent notamment un accès à des sièges de surplus, c'est-à-dire des sièges restés non réservés en raison d'une demande insuffisante ou de caractéristiques peu attrayantes, telles que de longues durées de voyage, des départs tardifs ou des itinéraires avec escales multiples. Toutefois, l'influence des différents facteurs sur la probabilité de réservation de ces sièges restent inexplorés, alors même qu'elle est essentielle pour affiner les stratégies commerciales et optimiser la rentabilité.

Ce mémoire a pour objectif d'estimer la probabilité de réservation des sièges de surplus sur divers marchés, en fonction de facteurs spécifiques. Il vise également à analyser l'effet de ces facteurs sur la probabilité de réservation, dans le but d'améliorer la gestion des réservations de surplus.

La méthodologie adoptée repose sur une préparation des données, permettant de traiter des informations pertinentes sur les réservations de surplus et les caractéristiques associées aux vols. Des techniques de classification ont été utilisés pour estimer la probabilité de réservation de surplus, et plusieurs algorithmes d'apprentissage automatique ont été comparés afin d'identifier le plus performant. Le modèle LightGBM s'est imposé comme le plus adapté à cette problématique. Trois stratégies de modélisation distinctes ont été mises en œuvre pour améliorer la précision des prédictions. Par ailleurs, l'effet des variables explicatives sur la probabilité de réservation a été exploré à l'aide de techniques de dépendance partielle et de mesures d'importance fondées sur la réduction de l'impureté de Gini.

La modélisation a permis d'atteindre des niveaux de rappel et de précision supérieurs à 80%, traduisant une forte capacité de prédiction. Des facteurs clés tels que la fenêtre de réservation, l'heure de départ et le mois de réservation se sont révélés particulièrement influents, chacun ayant un effet distinct sur la probabilité de réservation de surplus.

Ces résultats offrent des perspectives concrètes pour l'optimisation des programmes de fidélisation, en contribuant à une meilleure valorisation des sièges de surplus et à une rentabilité accrue.

ABSTRACT

Frequent flyer programs play a central role in brand management by enhancing brand recognition and fostering long-term relationships with customers. These programs offer access to surplus seats, i.e. seats that remain unreserved due to low demand or less attractive features, such as long travel durations, late departures, or itineraries with multiple layovers. However, the influence of various factors on the probability of reserving these seats remains unexplored, even though understanding it is essential for refining marketing strategies and optimizing profitability.

This dissertation aims to estimate the probability of surplus seat reservation across different markets, based on specific influencing factors. It also seeks to analyze the impact of these factors on the probability of surplus reservation in order to improve surplus reservation management.

The adopted methodology involved thorough data preparation to process relevant information on surplus reservations and associated flight characteristics. Classification techniques were used to estimate the probability of surplus reservation, and several machine learning algorithms were compared to identify the most effective one. LightGBM emerged as the most suitable model for this research. Three distinct modeling strategies were implemented to enhance predictive performance. In addition, the effects of explanatory variables on the probability of reserving surplus were analyzed using partial dependence techniques and feature importance measures based on Gini impurity reduction.

The modeling process achieved recall and precision levels exceeding 80%, indicating strong predictive capability. Key factors such as booking window, departure time, and the reservation month were found to be influential, each having a distinct effect on the probability of surplus reservation.

These findings offer practical insights for optimizing frequent flyer programs, contributing to better management of surplus seats and increased profitability.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

FFPs	Programmes de fidélisation
PNR	Enregistrement des noms de passagers
CRM	Gestion de la relation client
ASR	Réservation anticipée de sièges
EMSR	Revenu marginal prévu pour les sièges
SLR	Revue systématique de la littérature
AI	Intelligence artificielle
PRISMA	Éléments de rapport pour les revues systématiques et les méta-analyses
ACP	Analyse en composantes principales
kNN	k-plus proches voisin
GBMs	Machines de renforcement par gradient
AUC	Aire sous la courbe
GOSS	Échantillonnage unilatéral basé sur les gradients
EFB	Regroupement exclusif de caractéristiques
PDPs	Courbes de dépendance partielle
TRF	Taux de réservation final
NFA	Nombre final d'allocations
SMOTE	Technique de suréchantillonnage des minorités synthétiques
TP	Nombre de vrais positifs
FP	Nombre de faux positifs
FN	Nombre de faux négatifs
TN	Nombre de vrais négatifs

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

1.1 Contexte

Les programmes de fidélisation (FFPs) ont été introduits à la fin des années 1970 pour aider les compagnies aériennes à fidéliser leurs clients, à les encourager à revenir et à générer des flux de revenus supplémentaires, selon Wever [1]. Ce sont des outils de commercialisation qui renforcent l'engagement des clients. Ils fournissent également aux compagnies aériennes des données sur les préférences des clients et leurs habitudes de voyage, qui peuvent être utilisées pour adapter les services et les offres. En outre, ils peuvent aider les compagnies aériennes à gérer plus efficacement leurs allocations. En offrant des réservations de sièges et des surclassements sous forme de récompenses, les compagnies aériennes peuvent remplir des sièges qui pourraient autrement rester non réservés, optimisant ainsi la capacité et maximisant les revenus. Néanmoins, ils rencontrent un problème dans la réservation d'une catégorie de sièges en raison du manque d'informations pour mieux gérer leur réservation et adapter leurs stratégies de commercialisation. Avant d'expliquer ce problème, il est important de révéler comment ces programmes fonctionnent et à quel niveau du processus ce problème apparaît.

FFPs fournissent des cartes à leurs membres après leur inscription au programme et les membres peuvent gagner des points grâce à une variété d'activités – notamment en volant avec une compagnie aérienne, en dépensant avec des cartes de crédit co-marquées et en transférant des points à partir de plusieurs partenaires bancaires et hôteliers. Après avoir accumulé des points, il existe de nombreuses façons de les échanger, comme la réservation d'un vol ou d'un hôtel, la location d'une voiture ou l'achat d'une carte-cadeau ou de marchandises. À ce stade, après l'expérience de rédemption, deux décisions peuvent être prises : continuer à accumuler et à gagner des points, ce qui correspond à un réengagement, ou abandonner le programme. La figure 1.1 présente le cycle de vie des membres, ce qui est un moyen de comprendre les différentes dimensions et dynamiques qui se produisent dans le programme.

Afin de mieux comprendre ce cycle, chaque étape est expliquée plus en détail comme suit, en commençant par l'inscription des membres.

1.1.1 Inscription des membres

L'adhésion des membres repose principalement sur les aspirations individuelles des personnes. En effet, les motivations d'adhésion varient : certains membres souhaitent voyager pour le loisir, tandis que d'autres aspirent à profiter des avantages associés à leur statut de voyageur.

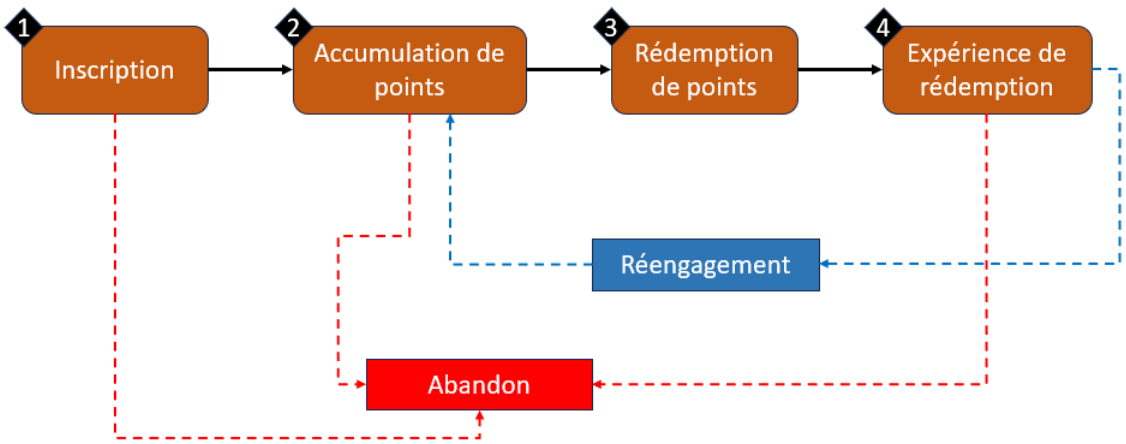


FIGURE 1.1 Cycle de vie des membres d'un programme de fidélisation

Ces différentes aspirations influencent directement la décision de rejoindre le programme. En outre, l'adhésion peut également être facilitée par un partenaire financier ou par un partenaire de vente au détail. Ces multiples points d'entrée permettent une flexibilité d'adhésion, adaptée aux préférences et aux besoins des futurs membres.

1.1.2 Accumulation de points

L'accumulation de points est un facteur clé qui détermine l'activation et l'engagement des nouveaux membres.

Les membres disposent d'options pour accumuler des points dans différents secteurs, comme le montre la figure 1.2 : le secteur financier, le secteur aérien, le secteur non aérien, et le commerce de détail, notamment grâce aux achats effectués auprès des détaillants ou via eStore du programme de fidélisation. Cette accumulation de points permet non seulement aux membres de bénéficier de récompenses, mais également de se soustraire à la politique d'expiration des points. Il existe également trois façons de gagner des points : soit directement, soit en convertissant des points d'un autre programme en points du programme de fidélisation, et enfin en achetant des produits d'affiliation.

Les programmes de conversion de points ont été des leviers pour attirer des membres dans le programme. L'utilisation des cartes de crédit co-marquées permet de gagner des points sur tous les achats, tandis que les achats effectués chez des partenaires offrent des points supplémentaires. De plus, les achats réalisés avec une carte de crédit co-marquée chez des partenaires permettent de doubler les points, ce qui incite les membres à utiliser ces cartes.

Les membres peuvent également accumuler davantage de points en convertissant ceux obtenus

Type de gain / Secteur	Gain direct	Ganger de l'argent avec les affiliés	Conversion
Secteur financier	TD CIBC AMERICAN EXPRESS		AMERICAN EXPRESS
Compagnies aériennes	AIR CANADA STAR ALLIANCE		
Compagnies non aériennes	AVIS MARRIOTT		Hilton MARRIOTT
Vente au détail	Uber Uber Eats LCBO	eStore	points

FIGURE 1.2 Secteurs et modes pour gagner des points du programme de fidélisation

dans d'autres programmes en points du programme de fidélisation ou en associant leurs comptes avec leurs marques préférées.

1.1.3 Rédemption de points

Après avoir accumulé des points pendant un certain temps, les membres envisagent comment maximiser les avantages du programme. En effet, la possibilité d'échanger des points est l'objectif de la majorité des membres lorsqu'ils adhèrent au programme – qu'il s'agisse d'atteindre un objectif de voyage ou d'acheter d'autres produits, les options de rédemption de points représentent un levier d'engagement. Les membres peuvent utiliser leurs points dans trois catégories de récompenses :

- Vols
- Hôtels et voitures de location
- Cartes-cadeaux et marchandises

Ils peuvent échanger leurs points contre des récompenses aériennes, non seulement avec la compagnie aérienne principale qui gère le programme de fidélisation, mais également avec d'autres partenaires aériens. Dans ce cas, le membre peut payer la totalité du vol en utilisant les points accumulés, mais il peut aussi ne payer qu'un pourcentage de la facturation. Il existe différentes méthodes de paiement, une combinaison de points et d'argent liquide dans la facture, comme suit :

- **Sans espèces** : signifie que la totalité du prix du billet a été payée avec des points et que les taxes ont également été payées avec des points.

— **Payer le total avec des points** : signifie que la totalité du prix du billet a été payée avec des points, mais que les taxes ont été payées en espèces.
— **Payer un pourcentage avec des points** : signifie que seulement un pourcentage du prix du billet a été payé avec des points et que le reste a été payé en espèces. Ce pourcentage varie selon les offres du programme de fidélisation.

L'équivalence des prix des vols en points dépend de la région, la classe de cabine et la distance parcourue. Différentes zones sont définies à cet effet, comme le montre par exemple, la figure 1.3, qui indique les limites spatiales entre l'Amérique du Nord, l'Amérique du Sud, l'Atlantique et le Pacifique. Ces régions comprennent différents pays et aéroports, et les vols peuvent être interrégionaux ou intrarégionaux.

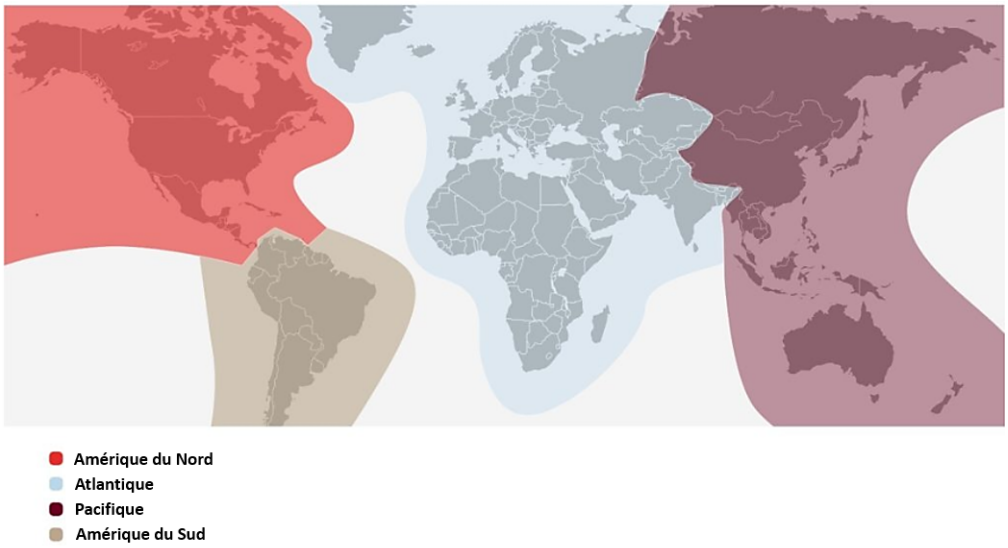


FIGURE 1.3 Zones de voyage

Le nombre de points requis varie en fonction de la classe de cabine : affaires, économique premium et économique. Il est à noter que plus le niveau de la classe est élevé, plus le nombre de points requis pour la conversion est important.

En outre, ils ont la possibilité de convertir leurs points en récompenses non aériennes, telles que des séjours à l'hôtel, des locations de voitures ou des articles de marchandise. Cette diversité de choix permet aux membres de bénéficier pleinement du programme, en fonction de leurs besoins et aspirations.

Ce processus renforce l'engagement des membres en leur offrant des bénéfices qui récompensent leur fidélité au programme.

1.1.4 Expérience de rédemption

Après une expérience de rédemption, les membres se posent des questions qui détermineront s'ils restent dans le programme ou s'ils décident de le quitter. Parmi ces questions figure la perception de la valeur des récompenses obtenues, que ce soient des vols ou autres. D'autres facteurs influencent également la décision de réinscription : la facilité du processus de rédemption, l'effort fourni et la rapidité d'obtention de la récompense. En ce qui concerne les vols, les membres évaluent l'éventail des options en termes de destinations et de dates disponibles.

Le but d'un programme de fidélisation est que les membres perçoivent la valeur réelle de leurs points et choisissent de s'engager à nouveau en commençant à accumuler des points pour atteindre leur prochain objectif de voyage.

Dans le cadre de cette recherche, l'accent est mis sur la troisième partie du cycle de vie des membres, qui est la rédemption de points, et plus particulièrement sur la rédemption par l'achat d'un billet d'avion. La problématique indique que certains sièges ne sont pas réservés. Ces sièges appelés "surplus" ont un faible volume de réservation par rapport au nombre de sièges alloués. La connaissance du moment où ces sièges sont plus probables d'être réservés permet de mieux gérer leur réservation. Cela nous amène à développer deux pistes de recherche, à savoir l'estimation de la probabilité de réservation des sièges de surplus et la détermination de l'effet des facteurs qui influencent cette probabilité. Pour ce faire, il convient d'abord d'expliquer certaines informations relatives aux réservations de sièges, telles que les détails qui caractérisent un billet, un vol et un siège.

1.1.5 Désignation d'un billet

Un billet d'avion est spécifié par un code unique pour chaque passager. Si un groupe voyage ensemble, une famille par exemple, ses codes de billets seront associés à un seul code d'enregistrement des noms de passagers (PNR) pour définir ce groupe. Un billet comprend une frontière pour un aller simple, ou deux pour un aller-retour. Chaque frontière peut avoir un ou plusieurs itinéraires, en fonction du nombre d'escales. Ces notions sont expliquées comme suit :

- **Code de PNR** : code associé à la réservation. Un PNR peut avoir plusieurs billets (un par passager). Par exemple, une famille voyage ensemble, elle a des billets différents pour chaque membre, mais un seul code PNR.
- **Code du billet** : code associé au voyage complet du passager. Un billet peut inclure plusieurs frontières, comme un aller simple correspondant à une frontière et un aller-retour à deux frontières.

— **Code du frontière** : correspond à un aller simple. Il inclut un ou plusieurs itinéraires. S'il n'y a pas d'escale, cela signifie qu'il y a un seul itinéraire ; une escale signifie qu'il y a deux itinéraires ; deux escales signifient qu'il y a trois itinéraires.

— **Code d'itinéraire** : correspond à un déplacement d'un point d'origine à un point de destination sans escale, qui est le niveau le plus granulaire.

La figure 1.4 illustre un exemple de code PNR pour un groupe de trois billets, soit trois personnes. Chaque personne possède un code de billet distinct. Ce voyage comprend un aller-retour, soit deux frontières. Pour chaque frontière, il y a une escale, ce qui signifie qu'il y a 2 itinéraires à l'aller et au retour.

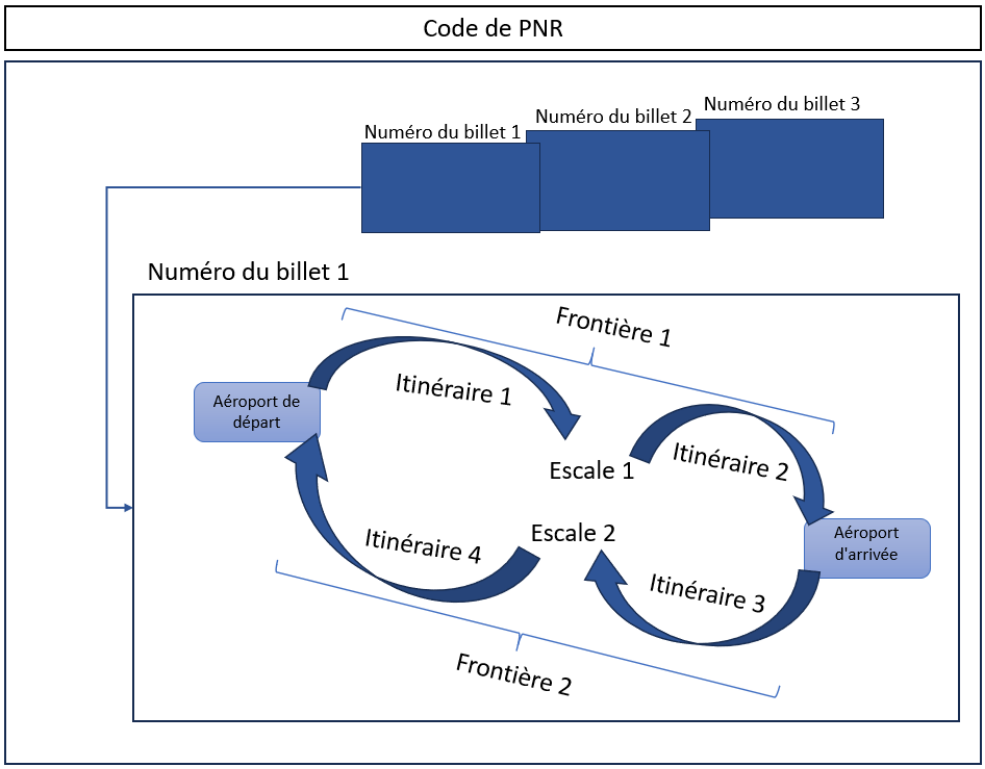


FIGURE 1.4 Explication des détails du billet

1.1.6 Définition d'un vol

Un vol est un voyage d'un point de départ à un point de destination sans escale à une date de départ précise. Il est défini par trois variables : itinéraire, date de départ et numéro de vol, expliquées comme suit :

— **Itinéraire** : direction entre l'aéroport du pays d'origine et l'aéroport de destination sans escale.

— **Date de départ** : date de début du vol en année, mois et jour.
— **Numéro de vol** : désignation de l'heure exacte du vol en heures et minutes. Pour un itinéraire et une date de départ donnés, on peut trouver différents vols partant à des heures différentes, chacun d'entre eux se voyant attribuer un numéro de vol pour les différencier.

La fixation de ces trois variables assure l'unicité du vol, signifiant qu'il n'y a qu'un seul vol avec ces caractéristiques.

1.1.7 Définition d'un siège

Un siège en vol est une place attribuée aux membres du programme en échange de points ou de miles gagnés lors de voyages précédents. Les sièges attribués à un vol peuvent également être appelés allocations. Un siège peut être assigné à trois types de classes en fonction du code de la cabine : affaires, économique premium et économique.

Le nombre de points requis pour échanger un vol libre dépend de divers facteurs tels que la ville d'origine, la ville de destination, le code de la cabine et les miles parcourus. Chaque vol peut être piloté soit par la compagnie aérienne principale, soit par une compagnie partenaire et, dans les deux cas, le vol peut avoir deux types de sièges : surplus ou dynamique, défini comme suit :

— **Surplus** : il s'agit d'un siège qui est resté non réservé par la compagnie aérienne et qui est donné à un programme de fidélisation pour être vendu à un prix réduit.
— **Dynamique** : il s'agit du siège présentant des caractéristiques favorables à la réservation, vendu au prix du marché.

La différence entre ces deux sièges est que le surplus présente une situation plus défavorable pour le voyage que le dynamique en termes de date de départ et d'arrivée, de destination, de compagnie aérienne, de région, de distance du voyage et de durée du vol, etc. Par ailleurs, le siège de surplus est celui qui présente des difficultés à réserver en raison de conditions de vol non adéquates pour le membre. Ces types de sièges sont définis a priori par la compagnie aérienne, ce qui signifie que le membre n'a pas accès à ces informations.

1.2 Explication de la problématique

La première étape pour expliquer la problématique consiste à définir l'axe de recherche. L'analyse se concentre seulement sur les données de la compagnie aérienne principale, en excluant celles des partenaires, car ces derniers représentent une minorité au sein de la base de données d'un programme de fidélisation.

La figure 1.5 montre l'axe spécifique sur lequel la recherche est assignée, qui est marqué en rouge, montrant les différents niveaux de siège et les compagnies aériennes avec lesquelles le programme de fidélisation est en contact.

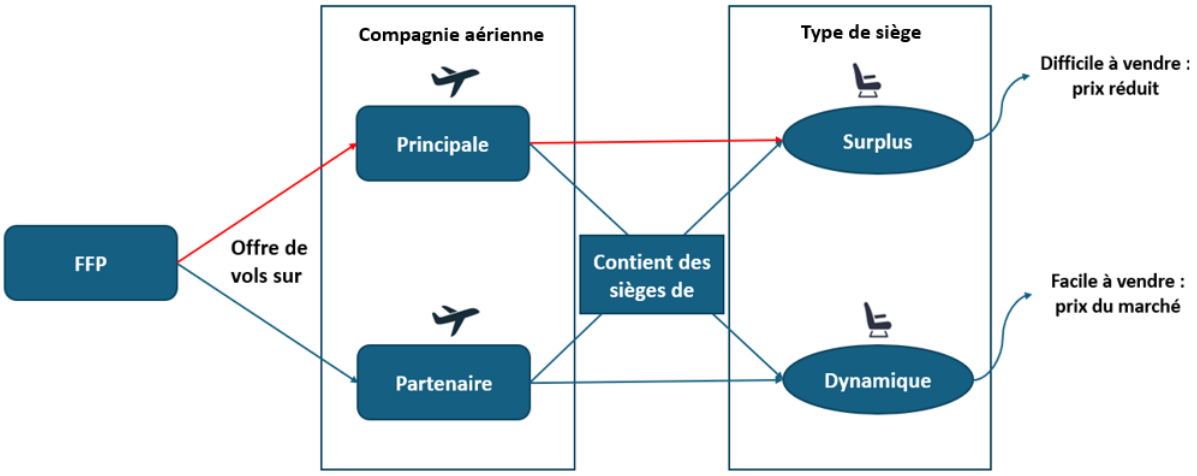


FIGURE 1.5 Axe de recherche

Le programme de fidélisation reçoit de la compagnie aérienne de sièges, soit de surplus ou dynamique, à réserver sur des vols spécifiques. Ce nombre de sièges est appelé **allocations** et est mis à jour quotidiennement, depuis un an avant la date de départ jusqu'à celle-ci. Les jours précédant la date de départ au cours desquels le nombre d'allocations change sont appelés **dates de publication**. FFPs cherche, par le moyen de son programme, à maximiser le pourcentage de réservation de ces sièges qui lui sont alloués. Dans cet axe de recherche, nous nous concentrons uniquement sur les informations relatives aux sièges de surplus et négligeons celles relatives aux sièges dynamiques.

La réservation de ces sièges dépend de différents facteurs. La figure 1.6 présente ces facteurs, tels que les caractéristiques personnelles, les informations relatives au billet et celles relatives aux réservations de surplus. Les informations personnelles, comme la fréquence de voyage d'un membre, son âge, les lieux les plus visités, etc., sont utilisées pour segmenter les clients et commercialiser les sièges par ciblage. En outre, les informations sur les billets auxquels ces sièges appartiennent, telles que le nombre de frontières, s'il s'agit d'aller-retour ou non, le nombre d'escales, le pays de l'escale si elle a lieu, le mode de paiement du billet, soit en espèces, en points ou une combinaison des deux, le nombre de points utilisés pour le paiement, etc., sont également considérées comme des facteurs influençant la réservation de sièges de surplus, puisqu'elles fournissent des informations sur les préférences des voyageurs et les activités des membres du programme.

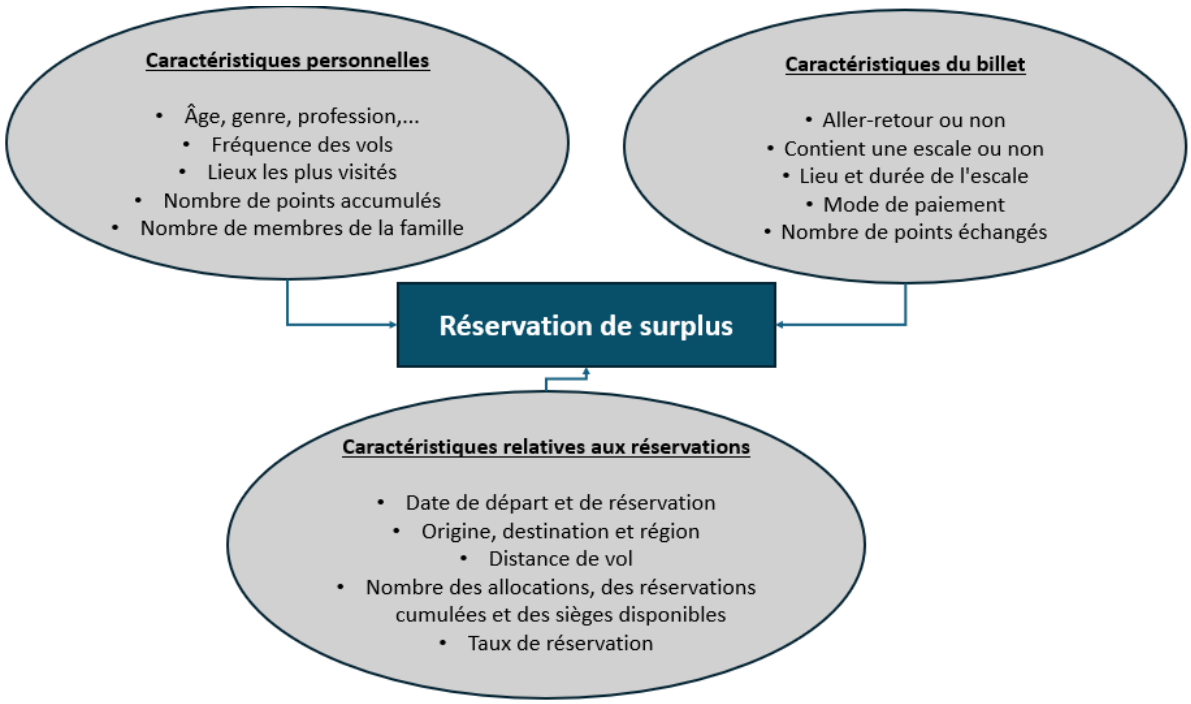


FIGURE 1.6 Facteurs relatifs à la réservation des sièges de surplus

Dans cette recherche, quelques caractéristiques ne sont pas prises en compte car elles fournissent des informations relatives à l'environnement externe et des données personnelles, en plus d'informations très granulaires telles que les caractéristiques du billet ou du siège, alors qu'il convient de se concentrer uniquement sur les informations relatives directement aux réservations de surplus, comme les détails concernant la date de réservation, le jour de la semaine et le mois de réservation, pour capter les effets saisonniers et les tendances temporelles. Cela permet également d'analyser l'existence de réservations de surplus en fin de semaine par rapport à d'autres jours et de prendre en compte les événements spécifiques ou les jours fériés selon le pays. Les mêmes connaissances peuvent être extraites de la date de départ. On considère aussi la distance du vol, l'origine, la destination, la région et la **fenêtre de réservation,** qui est le nombre de jours restants avant la date de départ.

Au niveau de la fenêtre de réservation pour un vol spécifique, comme indiqué dans la figure 1.7, il convient de considérer les informations sur le nombre d'allocations, qui représente le nombre de sièges alloués pour réservation à une date de publication, et le nombre cumulatif de réservations, qui représente le nombre total de sièges réservés jusqu'à une date de publication.

Deux autres informations ont également été introduites, à savoir le nombre de sièges disponibles et le taux de réservation. Le premier est défini comme la différence entre le nombre d'allocations et le nombre cumulatif de réservations. Le deuxième est le rapport entre le

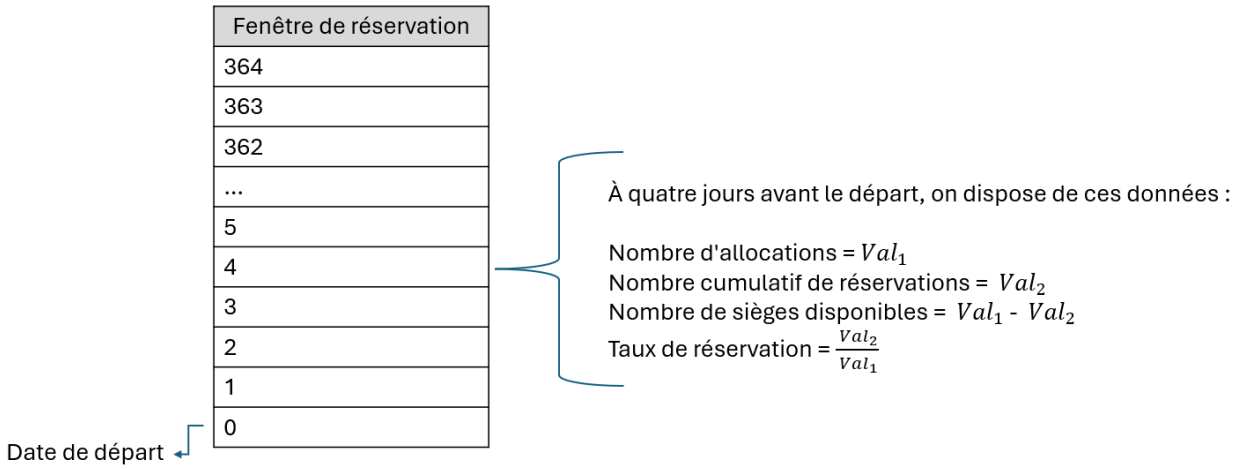


FIGURE 1.7 Informations au niveau de la fenêtre de réservation

nombre cumulatif de réservations et le nombre d'allocations.

Afin de comprendre l'évolution de ces variables au cours de la fenêtre de réservation et de mettre en évidence la difficulté de la réservation de surplus, on extrait les différents cas qui peuvent exister au cours de la fenêtre de réservation, en utilisant le tableau 1.1.

Deux cas ont été observés : le premier correspond à une absence d'allocations, et le second à une situation où des allocations sont disponibles, leur nombre pouvant varier en fonction de la fenêtre de réservation selon les ajustements effectués par le système d'optimisation de la compagnie aérienne. Dans les deux cas, l'estimation de la probabilité de réservation de surplus, en fonction du nombre d'allocations et des variables mentionnées précédemment, permet au programme de fidélisation d'identifier le moment pour demander plus d'allocations en vue de leur commercialisation et de leur réservation.

Le problème spécifique réside donc dans le manque d'informations sur la probabilité de réservation de surplus et l'effet de différents facteurs sur cette probabilité, comme la date de réservation, la fenêtre de réservation, la région, la distance du vol et l'heure de départ. Cela conduit à deux questions de recherche :

— Comment estimer la probabilité de réserver des sièges de surplus ?
— Quel est l'effet des facteurs responsables de la variation de cette probabilité ?

Cela mène à améliorer la gestion d'allocations et à affiner les stratégies de commercialisation, permettant d'augmenter le volume des réservations de surplus.

L'estimation de la probabilité de réservation de surplus est obtenue dans ce projet par la résolution d'un problème de classification. À cet égard, le but est de prédire, à l'aide d'un

Tableau 1.1 Variation d'allocations et réservations de surplus cumulées selon la fenêtre de réservation

Fenêtre de réservation	Allocations cumulées	Réservations cumulées
230	0	0
229	0	0
228	0	0
227	0	0
226	0	0
225	0	0
224	10	0
223	10	0
222	10	0
221	10	0
220	10	0
219	20	0
218	20	4
217	20	4
216	20	6
215	20	8
214	20	8

modèle de classification, l'état de réservation de surplus dans les sept jours à venir à chaque date de publication, c'est-à-dire s'il y a au moins un siège réservé dans les sept jours suivants. Une période de sept jours a été choisie, car elle est stratégiquement utilisée par la compagnie aérienne pour la gestion des allocations. Après avoir résolu ce problème de classification, le modèle appliqué permet d'extraire, pour chaque date de publication, la probabilité estimée de réserver des sièges de surplus. Une fois la première étape réalisée, l'étude se concentre sur l'analyse de l'influence de différents facteurs sur cette probabilité.

En résumé, cette recherche vise à mieux comprendre et prédire les dynamiques de réservation de sièges de surplus afin d'optimiser leur commercialisation dans le cadre des programmes de fidélisation.

1.3 Objectifs du mémoire

Ce mémoire vise à fournir aux programmes de fidélisation des connaissances qui permettent de réserver plus de sièges de surplus. Cela augmente la rentabilité de l'entreprise. Nous nous appuyons sur les technologies de science des données afin de trouver une solution qui présente des méthodes et techniques intelligentes pour comprendre les données de réservation des sièges

de surplus.

Pour cela, deux objectifs sont définis. Le premier objectif est d'estimer la probabilité de réserver des sièges de surplus. Cet objectif comprend différentes parties comme suit :

— Préparation des données relatives aux réservations de surplus et sélection des variables pour la modélisation.
— Spécification du processus de modélisation pour estimer la probabilité de réservation de surplus en utilisant la variable dépendante appropriée, les modèles et les stratégies de modélisation dans différentes régions.
— Comparaison entre les modèles utilisés en utilisant des métriques spécifiées. La comparaison se fait entre les modèles d'apprentissage automatique ou entre les performances de modélisation de différentes régions. Le modèle sélectionné est également comparé au modèle de référence.

Le deuxième objectif est d'étudier l'effet de différents facteurs comme le mois de réservation, la fenêtre de réservation, l'heure de départ et la région sur la probabilité de réservation des sièges de surplus. Pour ce faire, nous suivons les étapes suivantes :

— Comparer l'importance des variables explicatives.
— Étudier la dépendance entre les variables et l'influence indépendante de certaines variables sur la probabilité de réservation de surplus.

Après avoir atteint ces objectifs, le programme de fidélisation dispose de connaissances pour améliorer son processus de prise de décision.

1.4 Organisation du mémoire

Le mémoire est organisé en 5 chapitres. Le chapitre 2 concerne la revue de la littérature des études faites pour estimer la probabilité de réservation dans différents secteurs (aérien, hôtelier, restauration, etc.) et acquérir des connaissances sur les technologies utilisées et les limitations présentées dans leurs études. Une partie de la recherche est également consacrée à trouver des études abordant l'effet des facteurs sur la probabilité de réservation. Le chapitre 3 est consacré à l'explication de la méthodologie appliquée à la préparation des données. Le chapitre 4 explique les différentes stratégies de modélisation et la méthodologie utilisée pour démontrer l'effet des facteurs sur la probabilité de réservation de surplus. Enfin, le chapitre 5 présente l'application et les résultats des différentes stratégies expliquées dans le chapitre précédent, suivis de la conclusion générale de ce travail, ainsi que des limites et des perspectives d'amélioration.

CHAPITRE 2 REVUE DE LITTÉRATURE

Dans ce chapitre, une revue de la littérature est menée afin d'identifier les travaux en lien avec la problématique et les objectifs de ce projet. Ces objectifs s'inscrivent dans une démarche visant à améliorer la rentabilité de la compagnie aérienne et à renforcer la fidélisation des membres, puisque les sièges de surplus, vendus à tarif réduit, représentent une opportunité d'optimisation des revenus lorsqu'ils sont mieux réservés. Dans cette optique, la section 2.1 examine les stratégies mises en œuvre par les programmes de fidélisation pour accroître la rentabilité des compagnies aériennes et renforcer l'engagement des clients, tout en mettant en évidence les écarts entre ces approches et la problématique traitée dans ce projet. La section 2.2 propose ensuite une revue systématique des travaux portant spécifiquement sur cette problématique.

Les sections 2.3 et 2.4 s'attachent aux méthodes développées pour atteindre le premier objectif du projet : estimer la probabilité de réservation des sièges de surplus. Ces sections couvrent un large éventail de stratégies, des modèles mathématiques et séries chronologiques aux approches d'apprentissage automatique.

Enfin, la section 2.5 traite du second objectif, qui consiste à analyser l'effet de divers facteurs sur la probabilité de réservation. Elle met en évidence les variables explicatives identifiées dans les secteurs de l'hôtellerie, de la restauration et du transport aérien, et présente les méthodes mobilisées pour cette analyse, notamment les techniques de dépendance partielle et les mesures d'importance des variables basées sur l'impureté de Gini.

2.1 Stratégies pour améliorer la rentabilité des compagnies aériennes

Les programmes de fidélisation ont été identifiés comme des leviers stratégiques pour stimuler la rentabilité des compagnies aériennes, notamment parce qu'une part importante des revenus est générée par les membres de ces programmes. Orhun et al. [2] soulignent leur rôle dans cette dynamique, tandis que Lee-Anant [3] met en avant leur capacité à améliorer la rétention des clients et la performance financière globale des compagnies. Fard et al. [4] proposent la stratégie de surréservation, consistant à accepter plus de réservations que de sièges disponibles, afin de maximiser l'occupation et les revenus. Bazargan et al. [5] affirment que les FFPs représentent un retour sur investissement parmi les initiatives de gestion de la relation client (CRM). De plus, Wever [6] insiste sur le rôle des FFPs dans la capacité à fidéliser des voyageurs fréquents et internationaux, contribuant ainsi à la croissance des

revenus. L'élargissement du réseau de partenaires et l'offre de récompenses renforcent leur attractivité et encouragent la fidélité.

L'évolution des capacités d'analyse de données a permis aux FFPs de perfectionner leurs stratégies de commercialisation et d'optimiser la gestion des sièges. Park et al. [7] utilisent des techniques d'apprentissage profond pour analyser le risque de désengagement et améliorer la satisfaction des clients. Lohiya et al. [8] exploitent la science des données pour analyser la demande sur des paires de villes spécifiques et ajuster dynamiquement les prix. Ertuğrul et Şahin [9] appliquent le modèle de revenu marginal prévu pour les sièges (EMSR) pour allouer de manière optimale les sièges entre différentes classes tarifaires, permettant de concilier revenus anticipés et opportunités de dernière minute. Par ailleurs, Guerrini et al. [10] montrent comment l'intégration des données issues des moteurs de réservation en ligne dans un entrepôt permet aux compagnies de segmenter la clientèle et de personnaliser leurs offres. D'un point de vue comportemental, Bachmat et al. [11] démontrent que les passagers manifestent des préférences spécifiques pour certains sièges, comme éviter ceux du milieu, ce qui ouvre la voie à la mise en place de services de réservation anticipée (ASR). Wang et al. [12] développent un modèle statistique prédisant la probabilité de choix d'un siège selon des critères comme le prix, la compagnie ou l'anticipation. Enfin, Li et al. [13] insistent sur la possibilité pour les compagnies de différencier leurs produits, en ajustant les prix non seulement pour les sièges, mais aussi pour les services auxiliaires tels que les repas ou les bagages prioritaires.

Après avoir passé en revue ces stratégies axées sur la fidélisation et la rentabilité, cette recherche propose une autre approche : l'amélioration de la gestion des réservations de sièges de surplus. Contrairement aux approches reposant sur la récompense, la surréservation ou l'analyse des préférences clients, cette recherche explore un segment spécifique de la stratégie commerciale : la réservation optimisée de sièges de surplus, cédés à bas coût par la compagnie aérienne au programme de fidélisation. Ces sièges, bien qu'économiquement avantageux, sont souvent difficiles à réserver en raison d'un manque de compréhension des facteurs qui influencent la probabilité de leur réservation.

L'originalité de ce travail réside dans la proposition d'un cadre analytique permettant aux programmes de fidélisation d'anticiper la probabilité de réservation des sièges de surplus à partir d'un ensemble de facteurs explicatifs. Ces estimations constituent un outil décisionnel pour ajuster les allocations et orienter les actions de commercialisation. En analysant l'effet de ces facteurs, cette recherche contribue à une meilleure gestion des sièges de surplus et, in fine, à l'accroissement de la rentabilité.

2.2 Revue systématique de la littérature

Cette recherche adopte une revue systématique de la littérature (SLR), une méthode qui vise à fournir un résumé et objectif de l'état actuel des connaissances sur un sujet donné, comme le décrivent Pradana et al. [14]. Contrairement à une revue traditionnelle qui s'appuie souvent sur des études choisies de manière subjective, la SLR permet, selon Kraus et al. [15], d'identifier des conclusions globales, de répondre à des questions de recherche ciblées et de proposer des pistes pour des recherches futures. Sauer et Seuring [16] définissent les principales étapes d'une SLR : planification, exécution, rapport des résultats et sélection des études sur la base de critères d'inclusion et d'exclusion. Des outils tels que la base de données Compendex et le diagramme de flux PRISMA sont recommandés pour structurer le processus de sélection, comme le suggèrent Phillips et al. [17]. Enfin, Wang et al. [18] insistent sur l'importance de cette méthode pour garantir la précision, la transparence et la reproductibilité des revues.

Une stratégie en trois étapes a été appliquée : (1) une recherche approfondie dans les bases de données académiques, (2) l'application de critères de sélection, et (3) la réalisation d'une revue complète orientée par des questions de recherche spécifiques.

Peu d'études se concentrent explicitement sur l'estimation de la probabilité de réservation dans les secteurs du transport aérien, de l'hôtellerie ou de la restauration. Toutefois, des recherches sur l'annulation, également considérée comme un événement binaire, fournissent des informations pertinentes. En effet, ces deux événements — réservation et annulation — partagent des caractéristiques similaires, comme le moment de réservation, le prix, la distance du vol ou les préférences des clients. Les recherches sur les annulations aident à comprendre les motivations des clients, leur sensibilité au prix et aux caractéristiques du vol, et les tendances temporelles, qui influencent également les décisions de réservation. En outre, les méthodologies employées pour estimer les probabilités d'annulation, telles que l'analyse des séries chronologiques ou les modèles d'apprentissage automatique comme les arbres de décision et la régression logistique, sont adaptées pour estimer les probabilités de réservation. L'analyse conjointe de ces deux types d'événements permet donc d'enrichir la compréhension des comportements de réservation.

Afin de guider cette recherche, trois questions ont été formulées en relation avec les objectifs du projet. La première question concerne les différentes méthodes utilisées pour estimer la probabilité de réservation ou d'annulation, définie comme un événement binaire :

> **RQ1 : Comment estimer les probabilités de réservation ou d'annulation ?**

Dans un second temps, l'émergence de l'intelligence artificielle, en particulier de l'apprentis-

sage automatique, offre des outils pour la modélisation de la probabilité de réservation. Cela conduit à la formulation de la deuxième question :

> **RQ2 : Comment l'apprentissage automatique contribue-t-il à estimer la probabilité de réservation ou d'annulation ?**

Enfin, la troisième question se concentre sur l'identification et l'analyse des facteurs qui influencent cette probabilité. Plusieurs variables explicatives sont couramment mobilisées dans la littérature : distance du vol, origine et destination, date et fenêtre de réservation, nombre de sièges disponibles, etc. L'étude de leur impact repose sur l'utilisation d'outils d'analyse statistique et de visualisation, visant à capter les effets temporels, comportementaux ou géographiques.

> **RQ3 : Quels sont les facteurs qui influencent la probabilité de réservation, et comment peut-on capter leur effet ?**

Pour répondre à ces trois questions, la phase initiale de la SLR s'est appuyée sur deux bases de données académiques reconnues : *Engineering Village* et *Web of Science*. Le choix de ces plateformes vise à assurer une couverture étendue des publications pertinentes, issues de divers domaines liés à la modélisation de la demande, à la fidélisation et à l'optimisation des revenus. La figure 2.1 présente le diagramme PRISMA utilisé pour guider et documenter le processus de sélection des études.

Nous définissons ensuite le besoin d'information, qui fait référence aux exigences spécifiques ou aux critères qui guident la sélection et la récupération de la littérature au cours du processus de revue systématique.

Besoin d'information : Formuler une méthode d'estimation de la probabilité de réservation des sièges et investiguer l'effet des facteurs liés à la réservation à l'aide de l'intelligence artificielle.

La prochaine étape consiste à définir le plan conceptuel, qui constitue le cadre général orientant l'identification et la sélection de la littérature pertinente. Cette étape permet de circonscrire la portée de la revue, de préciser les concepts ou thématiques clés, ainsi que de fixer les critères d'inclusion et d'exclusion. Le plan conceptuel oriente également la sélection des mots-clés issus du besoin d'information et le choix des bases de données. Il sert de fondement au développement d'une stratégie de recherche systématique, tout en permettant une adaptation itérative fondée sur les premiers résultats et retours obtenus. Ce processus contribue à garantir la cohérence et l'efficacité de la revue. Le plan conceptuel de cette recherche est présenté dans le tableau 2.1.

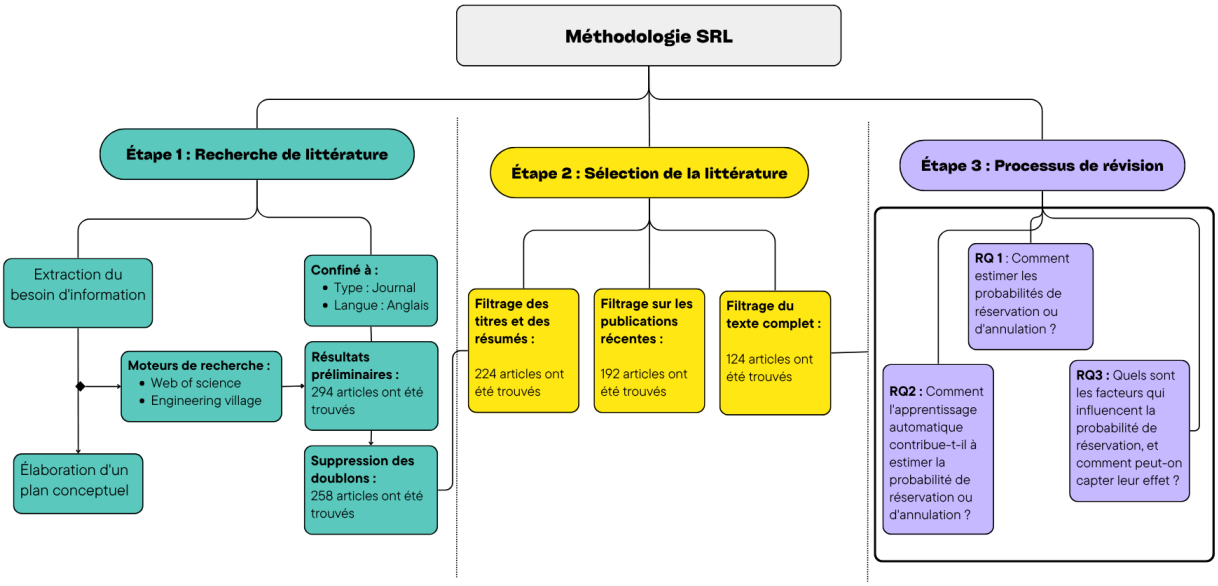


FIGURE 2.1 Organigramme PRISMA

Suite à la fusion des articles issus de chaque base de données, un total de 294 publications a été initialement identifié. Ce nombre a été réduit à 258 après l'élimination des doublons. Un processus d'examen a ensuite été mis en œuvre afin de sélectionner les publications les plus pertinentes pour cette recherche. Cette deuxième étape de la SLR s'est déroulée en trois phases : (1) un filtrage initial des titres et des résumés afin de retenir les articles en lien direct avec les questions de recherche ; (2) une exclusion des publications antérieures à dix ans ; et (3) une lecture complète des textes pour affiner la sélection, afin de garantir la mise à jour des connaissances utilisées.

À l'issue du filtrage des titres et résumés, 224 publications ont été considérées comme pertinentes pour une analyse complète. L'application du critère temporel a réduit ce nombre à 192. Une lecture approfondie a finalement permis de sélectionner 124 publications pour une analyse et une catégorisation détaillées.

Dans la dernière étape de la SLR, ces publications ont été classées selon les trois questions de recherche définies précédemment. Il est important de noter qu'une même publication peut contribuer à plusieurs questions, et que certains articles publiés il y a plus de dix ans ont été conservés en raison de leur importance méthodologique ou scientifique.

Tableau 2.1 Plan conceptuel

Thème	Mots clés
Transactionnalité	Booking* Cancel*
Probabilité	Probability Estimation Prediction
Ressource	Seat* Room* Inventor* Merchandise* Hotel* Flight*
Modélisation	Machine learning Data Data mining Mathematical model* Time series Statistical model*

La section suivante s'attache à répondre à la première question de recherche, portant sur les techniques utilisées pour estimer la probabilité de réservation ou d'annulation.

2.3 Estimation des probabilités de réservation et d'annulation : stratégies et techniques

L'estimation de la probabilité d'événements binaires, tels que la réservation ou l'annulation, revêt une importance dans l'optimisation de l'allocation des ressources dans divers secteurs, notamment la santé, l'hôtellerie, le transport et la vente au détail. Pour relever ce défi, les méthodologies employées incluent des modèles mathématiques, des modèles de choix discrets, des algorithmes d'apprentissage automatique et des modèles de séries chronologiques.

2.3.1 Modèles mathématiques

Dans le domaine de la santé, Kim et Giachetti [19] proposent un modèle mathématique stochastique de surréservation qui estime les probabilités de réservation en prenant en compte les distributions de probabilité associées aux absences (no-shows) — c'est-à-dire les clients ayant réservé mais ne se présentant pas — ainsi qu'aux arrivées sans rendez-vous (walk-ins). Leur approche repose sur des techniques d'estimation probabiliste, comme la distribu-

tion empirique des patients sans rendez-vous et l'approche naïve fondée uniquement sur les moyennes, permettant de dériver des probabilités conditionnelles liées à ces incertitudes. Ensuite, Carreras-García et al. [20] proposent un modèle fondé sur la programmation mixte en nombres entiers, intégrant les probabilités d'absence et d'annulation des patients ainsi que la variabilité des créneaux horaires, dans le but d'optimiser les revenus de la clinique.

Dans le secteur de l'aviation, Sulima [21] utilise une chaîne de Markov qui est un processus stochastique dans lequel la probabilité de passer à un état futur dépend uniquement de l'état actuel, et non des états précédents. Elle est appliquée pour modéliser le processus de réservation, en calculant la probabilité que les passagers réservent un vol en tenant compte des ventes de billets, des retours et des absences, et en utilisant la loi binomiale pour estimer les changements dans le nombre de billets au fil du temps.

Bien que ces techniques estiment les probabilités d'événements binaires, elles présentent des limitations. Selon Travin et al. [22], elles dépendent des hypothèses initiales et sont computationnellement complexes, limitant leur application. Enfin, ces approches présentent des difficultés à les interpréter.

2.3.2 Modèles de choix discrets

Dans le domaine de l'aviation, Shao et al. [23] ont introduit un modèle de durée en temps discret couplé à un modèle de choix multinomial afin d'estimer la probabilité de réservation anticipée de sièges (ASR) dans les compagnies aériennes. Ce modèle estime la probabilité qu'un passager effectue une réservation de siège à chaque jour avant le départ, en tenant compte de variables comme le prix relatif, le moment de la réservation et les préférences des passagers. Si une réservation est faite, un modèle de choix multinomial est appliqué pour prédire quel siège sera choisi, en fonction des caractéristiques des sièges et en intégrant des effets aléatoires pour capturer l'hétérogénéité entre sièges. Ce couplage permet de modéliser à la fois le moment de la décision et le choix précis du siège. Par ailleurs, Chiew et al. [24] proposent une approche bayésienne qui consiste à estimer un modèle de durée en temps discret reformulé en un modèle de choix binaire, où chaque jour le passager choisit d'annuler ou non. Les paramètres sont estimés par des méthodes de Monte Carlo par chaînes de Markov avec des coefficients aléatoires pour capter l'hétérogénéité individuelle. Cette approche permet de produire des probabilités d'annulation spécifiques à chaque individu, offrant à la fois la moyenne et la variance requises pour la gestion des revenus.

Cependant, les modèles de choix discrets présentent plusieurs limites. Ils reposent souvent sur des hypothèses fortes, telles que la linéarité des fonctions d'utilité, selon Xie [25]. De plus, ils peinent à capturer l'hétérogénéité non observée, la dynamique temporelle et les préférences

non linéaires, ce qui peut introduire des biais dans les estimations. Ces modèles deviennent également complexes à estimer en présence de jeux de données volumineux et peuvent souffrir de surapprentissage ou de difficultés de généralisation à de nouvelles données et à des scénarios inconnus, comme le soulignent Henley et al. [26].

2.3.3 Modèles des séries chronologiques

Les modèles de séries chronologiques sont utilisés pour analyser des données ordonnées dans le temps, permettant ainsi d'identifier des tendances évolutives, comme l'indique Percival [27]. Dans le contexte du covoiturage en libre-service, Müller et Bogenberger [28] ont appliqué ces modèles pour prévoir les comportements de réservation à court terme, permettant d'anticiper les périodes où une réservation est probable d'être effectuée. Par ailleurs, Tsai et Kimes [29] présentent un modèle de prédiction basé sur des cas de séries chronologiques pour prévoir les réservations de restaurants, démontrant son efficacité à prédire les motifs de réservation par rapport aux autres modèles de régression, abordant ainsi la probabilité des réservations à travers des techniques de prévision telles que la moyenne mobile intégrée autorégressive (ARIMA) qui est un modèle de séries temporelles qui combine l'autorégression, la différenciation et la moyenne mobile. Il est utilisé pour prédire des valeurs futures en tenant compte des dépendances passées et des tendances. Enfin, Jung et al. [30] ont évalué des modèles probabilistes de séries temporelles pour la prévision des ventes, et ont montré que des modèles alternatifs comme le perceptron multicouche et la régression linéaire surpassent ces approches en termes de performance, suggérant ainsi que les modèles temporels classiques ne sont pas toujours adaptés pour estimer les probabilités de réservation.

Bien que ces modèles soient utiles pour prévoir la demande, ils abordent généralement la question en termes de fréquence attendue plutôt qu'en termes de probabilité explicite d'occurrence d'une réservation. Les recherches se concentrent souvent sur l'identification des périodes favorables à la réservation, sans formuler directement la probabilité qu'un événement se produise.

L'estimation probabiliste par les séries chronologiques présente par ailleurs des limites. D'abord, elle repose sur les données historiques, c'est-à-dire sur les comportements et tendances observés dans le passé, ce qui peut compromettre la fiabilité des prévisions lorsque ces données ne reflètent pas les dynamiques futures, comme le soulignent Liu et al. [31]. Ensuite, ces modèles ont du mal à intégrer des facteurs contextuels ou externes importants pour la réservation, tels que les origines-destinations et les variations dans le nombre d'allocations. Enfin, ils reposent souvent sur l'hypothèse de stationnarité, qui ne s'accorde pas avec la réalité des données de réservation marquées par des effets saisonniers ou des tendances irrégulières.

Dans cette section, nous avons examiné diverses approches permettant d'estimer la probabilité d'événements tels que la réservation ou l'annulation. Des modèles mathématiques aux analyses de séries chronologiques, plusieurs méthodologies ont été explorées. Toutefois, peu d'études s'attachent à estimer directement la probabilité de réservation ; la majorité se focalise plutôt sur la prévision des volumes de réservation, principalement à des fins d'ajustement tarifaire. Les limites inhérentes à ces approches soulignent la nécessité de poursuivre les recherches afin d'améliorer à la fois la précision des estimations et l'interprétabilité des résultats. Cela permettrait de mieux comprendre les dynamiques de réservation dans des contextes variés. Dans ce cadre, nous avons exploré une autre voie : la modélisation par apprentissage automatique pour estimer la probabilité de réservation ou d'annulation. D'autres stratégies de modélisation sont également abordées dans la section suivante.

2.4 Techniques d'apprentissage automatique pour estimer la probabilité de réservation et d'annulation

Les algorithmes d'apprentissage automatique sont utilisés pour estimer la probabilité de réservation et d'annulation dans plusieurs secteurs, notamment l'hôtellerie et le transport. Ces estimations permettent aux décideurs d'anticiper la demande, d'optimiser l'allocation des ressources et d'améliorer la performance opérationnelle, comme l'indiquent Shah [32].

Dans le secteur hôtelier, les modèles prédictifs exploitent les données historiques pour anticiper les annulations et ajuster les stratégies de gestion. Rakesh et al. [33] montrent que les hôtels peuvent identifier les périodes où les annulations sont plus probables. Pour les plateformes d'hébergement de type "peer-to-peer", Afrianto et Wasesa [34] utilisent des modèles tels que la régression logistique, qui estime la probabilité d'un événement binaire à partir de variables explicatives, et les arbres de décision, qui segmentent les données en sous-groupes à l'aide de règles conditionnelles. Ces modèles permettent d'estimer la probabilité de réservation des logements.

Plusieurs approches d'apprentissage supervisé ont été testées, comme l'indiquent Kumar et Sharma [35], pour prédire les annulations, notamment les réseaux de neurones multicouches, qui modélisent des relations complexes à l'aide de couches successives de neurones artificiels, Adaptive Boosting (AdaBoost), une méthode d'ensemble qui combine plusieurs modèles faibles en les pondérant selon leurs performances, et Extreme Gradient Boosting (XGBoost), une version optimisée du gradient boosting, reconnue pour sa rapidité d'exécution. Juntong et Yin [36] proposent un modèle à deux niveaux combinant forêts aléatoires et analyse en composantes principales (ACP) afin d'améliorer la précision des prédictions. L'ACP est une technique de réduction de dimensionnalité qui transforme les variables corrélées en un en-

semble réduit de composantes principales non corrélées, tout en conservant l'essentiel de l'information. Cette étape permet de simplifier les données et de renforcer la performance des forêts aléatoires, un modèle d'ensemble basé sur de multiples arbres de décision. D'autres chercheurs, tels que Jishan et al. [37], intègrent les réseaux de neurones pour extraire les variables qui influencent la probabilité d'annulation. Tang [38] développe un système de classification combinant la méthode bayésienne naïve, les k-plus proches voisins et les forêts aléatoires, permettant ainsi aux hôtels de mieux estimer la probabilité de confirmation ou d'annulation des réservations.

Zhang et al. [39] s'intéressent à la gestion des incertitudes liées aux réservations. Leur modèle basé sur les forêts aléatoires prédit à la fois la demande et la probabilité d'annulation, permettant une tarification dynamique adaptée aux périodes de forte affluence. D'autres études, comme celle de Chen et al. [40], comparent plusieurs algorithmes, notamment le k plus proches voisins (KNN) et Categorical Boosting (CatBoost). KNN est un algorithme basé sur la similarité, qui prédit une observation en fonction des classes de ses voisins les plus proches, tandis que CatBoost est un algorithme de gradient boosting optimisé pour traiter efficacement les variables catégorielles. Les auteurs concluent que CatBoost offre les meilleures performances en termes de précision et de F1-score, grâce à sa capacité à réduire le surapprentissage.

Dans le transport, Sun et al. [41] conçoivent un réseau de neurones résiduel profond pour prédire la probabilité d'annulation des courses de covoiturage, en intégrant des variables comme la distance et le temps de trajet entre conducteur et passager. Dans le domaine aérien, Hopman et al. [42] utilisent XGBoost pour estimer les probabilités de réservation en incorporant des données externes telles que les prix concurrents ou les facteurs de sécurité. Cette approche permet de segmenter la clientèle et d'ajuster dynamiquement les stratégies de revenus.

L'ensemble de ces travaux montre que les modèles d'apprentissage automatique, bien qu'exigeants en calcul, offrent des gains significatifs en matière de performance prédictive selon Chen et al. [43]. Comme le soulignent Zhu et Chen [44], ces outils permettent une meilleure gestion des réservations, une efficacité accrue des opérations et une amélioration de la satisfaction client.

Les modèles de classification font partie des modèles d'apprentissage automatique examinés. Il propose différents types d'algorithmes pour modéliser des événements binaires tels que la réservation et l'annulation. Il extrait également les probabilités estimées de ces événements, ce qui en fait une solution à prendre en considération pour ce projet.

2.4.1 Importance des modèles de classification pour l'estimation des probabilités

Les techniques de classification sont couramment utilisées pour analyser les facteurs influençant le comportement de réservation, comme le soulignent Mpofu et al. [45]. Des algorithmes tels que les forêts aléatoires, les k-plus proches voisins, ou encore les arbres de décision renforcés par gradient (GBMs), sont mobilisés pour estimer la probabilité de réservation à partir de diverses variables explicatives.

Par ailleurs, Harris et Samorani [46] montrent que le choix du classificateur selon des métriques comme la perte logarithmique ou le score de Brier améliore la planification dans les systèmes de surréservation, comparativement à des indicateurs comme l'AUC.

Parmi les algorithmes de classification, les GBMs introduits par Friedman [47] présentent plusieurs avantages par rapport à des modèles plus classiques tels que la régression logistique ou les forêts aléatoires, notamment dans les cas de classification avec des données déséquilibrées. Contrairement aux modèles linéaires, qui peinent à capter des relations non linéaires ou des corrélations faibles entre les variables explicatives, les GBMs, grâce à leur structure arborescente, permettent de modéliser ces complexités, comme l'ont démontré Huang et Chen [48].

Comparés aux forêts aléatoires, les GBMs affichent une meilleure performance grâce à leur stratégie d'apprentissage séquentiel, chaque nouvel arbre corrige les erreurs de l'arbre précédent, ce qui améliore progressivement la précision du modèle, selon Osman et al. [49].

Les GBMs sont des modèles d'ensemble qui combinent plusieurs prédictions en une seule, renforçant ainsi la robustesse du modèle final. Des exemples populaires incluent XGBoost, LightGBM et CatBoost, tous capables de gérer des jeux de données déséquilibrés grâce à l'ajustement des poids d'échantillons et à l'utilisation de fonctions de perte personnalisées. Leur adaptabilité, conjuguée à leur capacité à fournir des mesures d'importance des variables explicatives, en fait des outils performants pour les tâches de classification selon Lu et Mazumder [50].

Pour cette recherche, l'algorithme LightGBM, développé par Microsoft, a été privilégié par rapport à XGBoost et CatBoost. Ce choix repose sur des critères de vitesse, d'efficacité mémoire et de scalabilité, comme le soutiennent Hosen et Amin [51]. LightGBM construit les arbres de décision de manière séquentielle via une descente de gradient, tout en adoptant une stratégie de croissance par feuilles. Son algorithme basé sur l'histogramme permet une consommation mémoire réduite, ce qui le rend particulièrement adapté aux jeux de données volumineux selon Qiu et al. [52].

En outre, LightGBM se distingue par sa capacité à effectuer un apprentissage parallèle et

distribué, ce qui favorise son utilisation à grande échelle suivant McCarty et al. [53]. Deux techniques spécifiques renforcent son efficacité :

- **Échantillonnage unilatéral basé sur les gradients (GOSS)** : conserve les échantillons avec les gradients les plus élevés afin d'optimiser l'apprentissage tout en réduisant la charge mémoire.
- **Regroupement exclusif de variables explicatives (EFB)** : regroupe de manière exclusive certaines variables explicatives pour réduire la dimensionnalité sans perte significative d'information selon Kwak et al. [54].

La suite de cette section examine en détail le fonctionnement de LightGBM et son utilisation pour estimer la probabilité de réservation de sièges de surplus.

2.4.2 Explication du processus de fonctionnement du modèle LightGBM

Le modèle de classification utilisé est LightGBM, qui est l'un des modèles GBMs. Son algorithme fonctionne de manière itérative pour améliorer les performances prédictives du modèle en minimisant une fonction de perte prédéfinie. Nous supposons que l'ensemble de données total contient n échantillons, répartis entre n_{train}, correspondant au nombre d'échantillons dédiés à l'apprentissage du modèle, et n_{test}, correspondant au nombre d'échantillons réservés au test, où $n = n_{train} + n_{test}$. On considère l'ensemble d'apprentissage $\{(x_i, y_i) \mid i = 1, 2, \ldots, n_{\text{train}}\}$,, où x_i désigne les valeurs des variables explicatives associées au i-ème échantillon, et $y_i \in \{0, 1\}$ représente la variable cible pour ce même échantillon. Par ailleurs, les prédicteurs x_j sont définis comme les variables explicatives liées au réservation de surplus.

La probabilité estimée de réservation de surplus pour i-ème échantillon, compte tenu des valeurs des variables explicatives dans x_i, est notée par $p_i = \Pr(y_i = 1 \mid x_i)$. Le fonctionnement de LightGBM est détaillé afin d'expliquer comment ces probabilités sont estimées. D'abord, il est nécessaire de définir certaines notions essentielles. Le logarithme des cotes de la probabilité de réservation de surplus (log-odds), noté z_i, est défini par :

$$z_i = \log\left(\frac{p_i}{1 - p_i}\right). \tag{2.1}$$

Le deuxième élément à définir est la fonction de perte d'entropie croisée binaire, notée L. Cette fonction mesure la divergence entre les étiquettes réelles de l'état de réservation, y_i, et les log-odds z_i. La fonction de perte initiale, en fonction de probabilité estimée p_i, est donnée par :

$$L = \sum_{i=1}^{n_{train}} L(y_i, p_i) = - \sum_{i=1}^{n_{train}} \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right]. \tag{2.2}$$

Par changement de variable via les log-odds, on obtient une formulation en fonction de z_i :

$$L = \sum_{i=1}^{n_{train}} L(y_i, z_i) = - \sum_{i=1}^{n_{train}} \left[y_i \log \left(\frac{e^{z_i}}{1 + e^{z_i}} \right) + (1 - y_i) \log \left(\frac{1}{1 + e^{z_i}} \right) \right]. \tag{2.3}$$

Le processus de construction du modèle débute par une initialisation à l'aide d'une valeur constante, puis se poursuit par des étapes successives d'ajustement des apprenants faibles.

Initialisation du modèle avec une valeur constante

La premier étape d'algorithme est d'initialiser le modèle avec une valeur constante qu'on l'appelle γ ou $z_i^{(0)}$. Cette valeur initiale représente l'estimation de base de log-odds.

$$z_i^{(0)} = \gamma = \arg\min_{\gamma} \sum_{i=1}^{n_{train}} L(y_i, \gamma). \tag{2.4}$$

La fonction de perte d'entropie croisée binaire pondérée $L(\gamma)$ est simplifiée en se basant sur l'équation 2.3 en préparation de la dérivation partielle. Cette fonction devient :

$$L(\gamma) = - \sum_{i=1}^{n_{train}} \left(y_i \gamma - \log(1 + e^{\gamma}) \right). \tag{2.5}$$

Afin d'obtenir γ pour lequel $L(\gamma)$ atteint son minimum, les dérivées partielles de la fonction de perte sont appliquées, et le gradient total de $L(\gamma)$ par rapport à γ est obtenu. Pour minimiser la fonction de perte, cette équation est mise en évidence :

$$\frac{\partial L(\gamma)}{\partial \gamma} = \sum_{i=1}^{n_{train}} \left(y_i - \frac{e^{\gamma}}{1 + e^{\gamma}} \right) = 0, \quad \text{où} \quad \frac{e^{\gamma}}{1 + e^{\gamma}} = \bar{y} \tag{2.6}$$

Cette équation montre que la différence entre les étiquettes réelles y_i et les probabilités estimées $\bar{y}$ doit être équilibrée pour tous les échantillons. La valeur de $\bar{y}$ est comme suit :

$$\bar{y} = \frac{1}{n_{train}} \sum_{i=1}^{n_{train}} y_i. \tag{2.7}$$

Cette équation montre que la valeur de $\bar{y}$ est telle que la probabilité moyenne estimée corresponde à la fréquence empirique des réservations.

Connaissant la valeur de $\bar{y}$ et sur la base de l'équation 2.19, la valeur initiale est égale.

$$z^{(0)} = \gamma = \log\left(\frac{\bar{y}}{1-\bar{y}}\right). \tag{2.8}$$

Raffinement du modèle

Cette étape présente des itérations dans laquelle le processus d'apprentissage ajoute de manière itérative de nouveaux apprenants faibles afin d'affiner le modèle. Pour $w = 1, 2, ..., M$, où M est le nombre total d'apprenants faibles, les étapes suivantes sont alors répétées :

1. Calcul des pseudo-résidus ou gradients négatifs :

$$r_{iw} = -\left[\frac{\partial L(y_i, z_i^{(w-1)})}{\partial z_i^{(w-1)}}\right]_{z_i^{(w-1)}}, \quad \text{pour } i = 1, 2, ..., n_{train} \tag{2.9}$$

   où $z_i^{(w-1)}$ est le log-odds de la probabilité estimée de réservation de surplus pour i-ème échantillon à l'itération $(w-1)$.
2. Ajustement de l'arbre de régression aux valeurs résiduelles r_{iw} et créer des régions terminales R_{qw} pour $q = 1, 2, ..., Q_w$. Les régions terminales sont les nœuds finaux ou les feuilles où un sous-ensemble particulier de points de données aboutit après avoir été divisé en fonction de différentes variables explicatives x_j.
3. Pour chaque feuille $q = 1, 2, ..., Q_w$ du nouvel arbre, le modèle calcule $\gamma_{q,w}$, qui est la valeur de sortie. La somme ne concerne que les points de données qui entrent dans la construction de cette feuille. En théorie, le modèle minimise localement la fonction de perte par dérivée partielle par rapport à γ pour obtenir la valeur de $\gamma_{q,w}$.

$$\gamma_{q,w} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{iq}} L\left(y_i, z_i^{(w-1)} + \gamma\right). \tag{2.10}$$

4. Mise à jour du modèle par actualiser $z_i^{(w)}$. Le taux d'apprentissage α évalue la contribution de chaque nouvel arbre ajouté au modèle. Il contrôle l'impact de chaque arbre individuel sur le modèle global. Des valeurs plus faibles de α se traduisent par des mises à jour plus petites, ce qui peut ralentir le processus d'apprentissage, mais conduit à un modèle plus robuste en réduisant le surajustement.

$$z_i^{(w)} = z_i^{(w-1)} + \alpha \sum_{q=1}^{Q_w} \gamma_{q,w} I\left(x_i \in R_{q,w}\right). \tag{2.11}$$

Estimation de la probabilité d'une réservation de surplus et prédiction de l'état de rerservation

Après M itérations, le modèle LightGBM conserve l'ensemble des arbres construits, chacun contribuant à l'estimation finale par une mise à jour additive dans l'espace des log-odds. Pour chaque échantillon x_i, ces mises à jour successives permettent d'obtenir la valeur finale de log-odds, notée $z_i^{(M)}$. Celle-ci s'exprime comme la somme cumulative des contributions individuelle de chaque arbre :

$$z_i^{(M)} = z_i^{(0)} + \alpha \sum_{w=1}^{M} f_w(x_i), \tag{2.12}$$

où $f_w(x_i)$ désigne la contribution en log-odds de l'arbre w-ième pour l'échantillon x_i, définie comme :

$$f_w(x_i) = \sum_{q=1}^{Q_w} \gamma_{q,w} \cdot I(x_i \in R_{q,w}), \tag{2.13}$$

Dans le cas particulier où aucun rétrécissement d'apprentissage n'est appliqué, c'est-à-dire $\alpha = 1$, et si la distribution de la variable cible est équilibrée, $\bar{y} = 0.5$ et $z_i^{(0)} = 0$, la valeur finale de log-odds devient simplement :

$$z_i^{(M)} = \sum_{w=1}^{M} f_w(x_i). \tag{2.14}$$

Ainsi, la probabilité estimée de réservation de surplus dans les sept jours à venir est alors obtenue en appliquant la fonction sigmoïde au log-odds final :

$$p_i = \Pr(y_i = 1 \mid x_i) = \frac{1}{1 + e^{-z_i^{(M)}}}. \tag{2.15}$$

Ce calcul est réalisé de manière identique, que l'échantillon appartienne à l'ensemble d'apprentissage, de validation ou de test. La prédiction repose uniquement sur la somme des contributions des arbres appris.

Durant l'apprentissage, des techniques de régularisation sont appliquées, telles que la réduction du taux d'apprentissage, la limitation de la profondeur des arbres, et le sous-échantillonnage, afin de prévenir le surajustement. Le processus de renforcement est interrompu après un nombre fixé d'itérations ou dès que les performances cessent de s'améliorer, selon un critère d'arrêt anticipé.

L'état de réservation de surplus est ensuite déterminé à partir de la probabilité estimée p_i, en comparant celle-ci à un seuil s :

$$\hat{y}_i = \begin{cases} 1 & \text{if } p_i \geq s, \\ 0 & \text{if } p_i < s. \end{cases} \tag{2.16}$$

Le seuil s est fixé par défaut à 0,5, mais il peut être ajusté pour optimiser une métrique de performance, telle que le F1-score, en particulier en présence d'un déséquilibre entre les classes.

Ainsi, l'algorithme LightGBM offre une méthode efficace pour estimer la probabilité de réservation de surplus, tout en s'adaptant à la nature déséquilibrée des données et en conservant une forte capacité de généralisation.

Enfin, d'autres modèles de classification ont été considérés à des fins de comparaison pour l'estimation de la probabilité de réservation ou d'annulation, LightGBM ayant été retenu pour ses performances supérieures.

2.4.3 Estimation des probabilités de réservation de surplus à l'aide de modèles de classification

Dans un modèle de classification, la probabilité de réservation de surplus correspond au degré de confiance attribué par le modèle à la survenue d'une réservation pour un vol donné, comme l'indiquent Hartini et al. [55]. Il s'agit d'une quantification de la certitude du modèle à l'égard de sa prédiction. Par ailleurs, la possibilité de comparer les probabilités issues de différents classificateurs offre un levier pour évaluer la qualité de l'estimation probabiliste de chaque modèle, selon Maxwell et al. [56].

L'analyse des divergences entre les probabilités estimées par différents classificateurs peut également être exploitée pour améliorer la performance globale du système, notamment à travers des techniques d'ensemble ou par l'ajustement des seuils de classification, comme le proposent Fang et Zhang [57]. Globalement, les probabilités de réservation de surplus jouent un rôle central dans la prise de décision, l'évaluation comparative des modèles et l'optimisation des tâches de classification liées à l'état de réservation.

Afin de comparer les approches probabilistes adoptées par différents modèles de classification, le tableau 2.2 synthétise les formules de sortie, ainsi que leurs principaux avantages et limitations.

2.5 Effet des facteurs sur la probabilité de réservation

Après avoir présenté les différentes techniques utilisées pour estimer la probabilité de réservation de surplus, cette section se concentre sur l'analyse des facteurs susceptibles d'influencer cette probabilité. De nombreuses études ont été menées pour identifier les variables déterminantes dans le processus de réservation, en particulier dans le secteur aérien, où les préférences des passagers, les dynamiques de marché et les considérations opérationnelles jouent un rôle central.

Selon Fuyane et al. [58], les attributs personnels, les caractéristiques du voyage et les conditions du marché sont des éléments clés dans la formation des préférences de réservation, notamment en ce qui concerne les horaires de départ ou d'arrivée. Banerji et al. [59] soulignent l'importance de la qualité du service client et de la ponctualité, notamment dans les économies émergentes, où ces facteurs influencent fortement la perception et la fidélité des passagers. De leur côté, Kobaszyńska-Twardowska et al. [60] insistent sur le rôle des éléments externes, tels que les conditions météorologiques ou les crises économiques, dans l'efficacité du système de transport aérien.

D'autres auteurs, comme Aleksić et al. [61], mettent en évidence l'impact de la fréquence des vols, de la distance, de la qualité du service et du prix, en soulignant que les préférences varient selon les segments de voyageurs. Ullah et al. [62] montrent que la pandémie de COVID-19 a modifié la sensibilité au prix et les comportements de consommation, notamment en ce qui concerne les stratégies promotionnelles.

TABLEAU 2.2 Probabilités extraites des modèles de classification

Modèle	Description	Formule	Avantages	Limitations
Régression Logistique	Elle estime la probabilité qu'un échantillon donné appartienne à une classe indiquant une réservation ou non en utilisant la fonction logistique (sigmoïde).	$\Pr(y_i = 1 \mid x_i) = \frac{1}{1+e^{-(b_1 x_i + b_2)}}$, — $b_1 x_i + b_2$ est la combinaison linéaire des variables explicatives x_i avec les poids b_1 et le biais b_2	— Interprétation des coefficients. — Bonne performance avec des données linéaires. — Peu coûteux en termes de calcul.	— Hypothèse de linéarité restrictive. — Sensible aux valeurs aberrantes. — Peu performant avec des relations non linéaires complexes.
Arbres de Décision	Ils ne fournissent pas directement des probabilités, mais infèrent des probabilités basées sur la distribution des classes des échantillons d'apprentissage, qu'il y ait ou non une réservation, dans le nœud feuille atteint par un échantillon particulier lors de l'inférence.	$\Pr(y_i = 1 \mid \text{nœud feuille}) = \frac{N_{1f}}{N_f}$, — $\Pr(y_i = 1 \mid \text{nœud feuille})$ représente la probabilité de réservation de surplus dans le nœud feuille. — N_{1f} représente le nombre d'échantillons de classe indiquant la réservation de surplus dans le nœud feuille. — N_f représente le nombre total d'échantillons dans le nœud feuille.	— Convient aux relations non linéaires. — Pas besoin de normaliser les données.	— Sensible au surapprentissage. — Performances limitées avec des données bruitées.
Forêt Aléatoire	Ils agrègent les probabilités en se basant sur la distribution des classes des échantillons d'apprentissage dans le nœud feuille atteint par un échantillon donné lors de la prédiction, en combinant les résultats de plusieurs arbres de décision pendant le test. Les arbres sont entraînés de façon indépendante, contrairement à LightGBM qui procède de manière séquentielle.	$\Pr(y_i = 1 \mid x_i) = \frac{1}{N_d} \sum_{d=1}^{N_d} Pr_1^{(d)}(y_i = 1 \mid x_i)$, — $\Pr(y_i = 1)$ est la probabilité d'appartenir à la classe indiquant la réservation de surplus. — N_d est le nombre total d'arbres dans la forêt. — $\Pr_1^{(d)}(y_i = 1 \mid x_i)$ est la probabilité estimée de réservation de surplus par le $d^{\text{ème}}$ arbre.	— Robuste contre le surapprentissage. — Gère les relations non linéaires. — Performant avec des données bruitées.	— Moins interprétable que les arbres uniques. — Nécessite plus de temps et de mémoire. — Moins efficace sur des données avec peu de variables explicatives.
k-plus proches voisins (kNN)	Il assigne la classe en fonction du vote majoritaire parmi ses k-plus proches voisins. Il dérive des probabilités en fonction de la fraction des voisins de chaque classe.	$Pr(y_i = 1 \mid x_i) = \frac{k_1}{k}$, — $\Pr(y_i = 1 \mid x_i)$ représente la probabilité que le point de données x_i appartienne à la classe indiquant la réservation de surplus. — k_1 représente le nombre de voisins appartenant à la classe indiquant la réservation de surplus parmi les k plus proches voisins de l'échantillon x_i. — k est le nombre total de voisins considérés.	— Simple à implémenter. — Performant sur des données bien séparées. — Aucun besoin d'entraînement explicite.	— Lent avec de grandes bases de données. — Sensible au bruit et à l'échelle des données. — Nécessite un choix de k.
LightGBM	Il estime les probabilités en moyennant les prédictions des arbres et en appliquant des transformations comme la sigmoïde. Il utilise des arbres de décision comme apprenants de base avec une croissance orientée par feuille, ce qui améliore l'efficacité et la précision. Cette méthode permet de construire des modèles plus rapidement et avec moins de mémoire, tout en offrant des probabilités fiables.	$\Pr(y_i = 1 \mid x_i) = \frac{1}{1+e^{-z_i^{(M)}}}$, — $z_i^{(M)}$ est le log-odds final de la probabilité estimée de réservation de surplus pour i-ème échantillon. — M indique la derniére itération dans le raffinement du modèle.	— Rapide et économe en mémoire. — Performant pour des ensembles de données larges. — Gère le déséquilibre de classes.	— Moins interprétable. — Nécessite un ajustement des hyperparamètres.

Des travaux portant sur les techniques de prévision, comme ceux de Fan et al. [63], ont montré que la régression par support vectoriel permet de capturer les tendances saisonnières et les fluctuations non linéaires de la demande de vol. Lurkin et al. [64] ont quant à eux mis en lumière l'efficacité des stratégies de tarification dynamique, en particulier à l'approche de la date de départ, pour adapter l'offre aux fluctuations de la demande.

L'influence des dimensions temporelles a également été analysée. Dutta et Santra [65] soulignent que l'écart entre la date de réservation et la date de départ a un impact sur les tarifs, tandis que Brey et Walker [66] montrent que l'heure de la journée affecte la demande selon le profil du voyageur. Chang et al. [67] remettent en cause l'idée d'une régularité dans les réservations selon les jours de la semaine, en révélant des effets comparables à ceux observés sur les marchés financiers, tout en précisant que ces effets varient selon les marchés étudiés.

Les facteurs liés à l'origine et à la destination sont essentiels dans la dynamique des réservations aériennes. Sandhu et Klabjan [68] proposent une approche qui tient compte des revenus des passagers dans la modélisation des itinéraires, tandis que Pölt [69] démontre l'efficacité des mécanismes de contrôle axés sur les itinéraires pour améliorer la gestion des revenus. Birolini et al. [70] soulignent quant à eux l'importance de la substituabilité des destinations et de la transparence des produits dans l'analyse du comportement de réservation.

Dans le cadre de cette recherche, nous nous concentrons sur un sous-ensemble spécifique de facteurs : la date de départ, la date de réservation, la fenêtre de réservation, l'itinéraire ainsi que le nombre de sièges de surplus alloués. Alors que la plupart des travaux existants se contentent d'évaluer l'impact de ces facteurs sur la décision binaire de réservation, notre étude cherche à analyser plus finement leur influence sur la probabilité estimée de réservation, un angle d'approche encore peu traité dans la littérature.

Pour ce faire, nous adoptons deux techniques complémentaires. D'une part, l'importance des variables est évaluée à l'aide de l'impureté de Gini, issue du modèle LightGBM, afin d'identifier les variables explicatives qui jouent un rôle déterminant dans la prédiction de la réservation de sièges de surplus. D'autre part, l'analyse de la dépendance partielle permet de visualiser l'effet marginal de chaque variable explicative sur la probabilité de réservation, en maintenant constantes les autres variables du modèle. Ces deux approches permettent non seulement de mieux comprendre les dynamiques sous-jacentes, mais aussi d'appuyer les décisions stratégiques des compagnies aériennes en matière de gestion des sièges de surplus. Elles sont détaillées dans les sections suivantes.

2.5.1 Importance des variables à l'aide de l'impureté de Gini

Dans cette section, la méthode d'évaluation de l'importance des variables explicatives basée sur l'impureté de Gini est utilisée pour mesurer l'impact de chaque variable sur les prédictions de l'état de réservation. L'importance des variables explicatives permet de quantifier quelles variables influencent le plus la capacité du modèle à distinguer entre la présence ou l'absence de réservation de surplus.

Impureté de Gini

L'impureté de Gini est une mesure qui quantifie la pureté des noeuds dans un arbre de décision du modèle LightGBM. Un nœud est dit "pur" s'il contient des échantillons d'une seule classe, c'est-à-dire uniquement des échantillons où une réservation de surplus est présente ou absente d'après Yuan et al. [71]. L'impureté de Gini évalue dans quelle mesure un ensemble de données est mélangé en termes de classes, et elle est utilisée lors de la construction des arbres de décision pour déterminer les séparations possibles selon Dunne et al. [72].

L'impureté de Gini pour un ensemble de données est définie par la formule suivante :

$$I_G = 1 - \sum_{a=1}^{A} t_a^2. \tag{2.17}$$

Où :

- t_a est la proportion d'échantillons de la classe a dans le noeud, c'est-à-dire la proportion de cas indiquant l'existence ou l'absence de réservation de surplus.
- A est le nombre total de classes, dans notre cas, $A = 2$, car il s'agit de classification binaire.

L'impureté de Gini prend une valeur comprise entre 0 (noeud pur) et 0,5 (noeud avec des proportions égales entre les deux classes). Lorsqu'un division est effectué dans un arbre de décision, il vise à réduire l'impureté de Gini, en séparant les données de manière à ce que les noeuds résultants soient plus purs.

Calcul de l'importance des variables explicatives

L'importance des variables explicatives relatives à la réservation des sièges de surplus dans LightGBM est mesurée par la contribution de chaque variable à la réduction de l'impureté de Gini au cours de la construction de l'arbre. Chaque fois qu'une division est effectué sur une variable dans l'arbre, l'impureté de Gini des noeuds descendants est réduite par rapport au noeud parent selon Scornet [73]. L'importance d'une variable explicative est alors définie

comme la somme des réductions d'impureté qu'elle génère à travers tous les divisions où elle est utilisée dans l'ensemble des arbres du modèle.

La réduction d'impureté de Gini pour une division particulier peut être calculée par :

$$\Delta I_G = I_G^{\text{parent}} - \left(\frac{N_L}{N} I_G^{\text{gauche}} + \frac{N_R}{N} I_G^{\text{droit}} \right), \tag{2.18}$$

Où :

- I_G^{parent} est l'impureté de Gini avant la division.
- I_G^{gauche} et I_G^{droit} sont les impuretés des noeuds fils gauche et droit respectivement après la division.
- N_L et N_R sont les nombres d'échantillons dans les noeuds gauche et droit, et N est le nombre d'échantillons dans le noeud parent.

Ainsi, pour chaque division où une variable est utilisée, sa contribution à la réduction de l'impureté de Gini est calculée. L'importance totale de la variable explicative est la somme de ces contributions sur l'ensemble des arbres du modèle. L'importance de la variable explicative est normalisée en pourcentage de la contribution à la réduction totale des impuretés. Plus une variable contribue à réduire l'impureté de Gini, plus elle est considérée comme importante. Dans cette recherche, nous examinons la contribution de diverses variables explicatives telles que le mois de la réservation, le jour de la semaine du départ et de la réservation, l'heure de départ du vol et la distance du vol, etc., dans la prédiction de l'état de réservation de surplus.

2.5.2 Technique de dépendance partielle

La technique de dépendance partielle est utilisée pour analyser, en moyenne, l'effet d'une variable explicative sur la probabilité estimée par un modèle, tout en maintenant constantes les autres variables. Elle est appliquée pour interpréter les modèles d'apprentissage automatique, en particulier ceux fondés sur des arbres de décision, comme le modèle LightGBM.

Pour une variable explicative x_j, l'objectif est d'évaluer l'impact de ses différentes valeurs sur la sortie du modèle, en moyennant sur les autres variables. Cette méthode a été proposée par Friedman et Popescu [74].

Pour calculer la dépendance partielle associée à une variable explicative x_j, les étapes suivantes sont appliquées :

1. **Sélection de la variable** : Identification de la variable explicative x_j à étudier.
2. **Définition des valeurs** : Construction d'une grille de valeurs que la variable x_j peut prendre, notée $\{v_1, v_2, \ldots, v_U\}$.

3. **Modification des données** : Pour chaque valeur v_u de la grille :
   — Création d'une copie de l'ensemble d'apprentissage.
   — Remplacement de la variable x_j par la valeur choisie v_u dans toutes les échantillons.
   — Conservation des autres variables x_{c_j} inchangées.
4. **Estimation** : Génération, à l'aide du modèle entraîné, des probabilités de la classe positive $p_i(x_j = v_u, x_{c_j})$ pour chaque échantillon modifié, avec les autres variables x_{c_j} inchangées.
5. **Transformation logit** : Afin de faciliter l'interprétation, une application de la fonction logit à chaque probabilité estimée a été effectuée, en vue d'une transformation en log-odds. Cette opération permet d'étendre l'intervalle de sortie de $-\infty$ à $+\infty$, contrairement aux probabilités bornées entre 0 et 1.
6. **Marginalisation** : Pour chaque valeur v_u, la moyenne des log-odds est calculée sur l'ensemble des échantillons. La dépendance partielle est alors exprimée par :

$$PD(x_j = v_u) = \frac{1}{n_{train}} \sum_{i=1}^{n_{train}} \log\left(\frac{p_i(x_j = v_u, x_{c_j})}{1 - p_i(x_j = v_u, x_{c_j})}\right). \tag{2.19}$$

Dans les graphiques de dépendance partielle (PDPs), la dépendance partielle est centrée et représentée sur l'axe vertical pour chaque valeur de la variable explicative x_j, afin d'illustrer de combien les log-odds sont supérieurs ou inférieurs à ceux de la probabilité moyenne. L'influence exercée par une variable sur la probabilité de réservation de surplus est ainsi quantifiée.

Lorsque les valeurs de la dépendance partielle sont proches de zéro, une influence marginale de la variable considérée sur la probabilité de réservation est indiquée. Plus ces valeurs s'éloignent de zéro, plus l'effet exercé par la variable est important. Des valeurs positives plus élevées signifient un effet positif important, augmentant la probabilité de réservation de surplus. En outre, des valeurs négatives plus élevées traduisent un effet négatif important, diminuant la probabilité de réservation de surplus. L'effet diminuant ou augmentant la probabilité de réservation est relatif à la probabilité moyenne. Si les valeurs de la dépendance partielle demeurent nulles pour l'ensemble des valeurs de x_j, aucun effet mesurable n'est observé pour cette variable sur la probabilité estimée.

CHAPITRE 3 PRÉPARATION DES DONNÉES

Dans cette recherche, Air Canada est considéré en tant que compagnie aérienne et Aéroplan comme son programme de fidélisation pour l'aider à résoudre les problèmes liés aux sièges de surplus. Le premier objectif est d'estimer la probabilité de réservation de surplus dans les sept prochains jours pour différents vols. Pour ce faire, un modèle d'apprentissage automatique supervisé est utilisé, en particulier la classification, pour prédire l'état de réservation de surplus dans les sept jours à venir. Ce dernier est une variable binaire qui indique à chaque date de publication s'il y a ou non au moins une réservation de surplus dans les sept prochains jours, 1 en cas de présence et 0 en cas d'absence. Cette variable est désignée par "état de réservation". Après modélisation, ce modèle est utilisé pour extraire la probabilité estimée de réservation de surplus dans les sept jours à venir. Le deuxième objectif est de comprendre l'effet des facteurs qui influencent cette probabilité en utilisant des techniques de dépendance partielle et l'importance des variables à l'aide de l'impureté de Gini.

Ce chapitre commence par expliquer la méthodologie utilisée pour la préparation des données dans la section 3.1. Ensuite, une analyse exploratoire des difficultés liées aux réservations de surplus a été faite à la section 3.2. Les données disponibles dans l'étude de cas Air Canada sont présentées à la section 3.3. Enfin, dans la section 3.4, l'application de la méthodologie de préparation des données est abordée.

3.1 Méthodologie de préparation des données

La méthodologie présentée dans la figure 3.1 vise à préparer les données pour modéliser l'état de réservation. Elle est appliquée à l'aide de deux ensembles de données D_{2023} et D_{2022} définis dans la section 3.3, concernant les départs de vols en 2023 et 2022 respectivement. Les tendances de réservation diffèrent entre ces deux années, et les caractéristiques des vols varient d'une région à l'autre.

La première étape est consacrée à l'extraction des variables présentes dans les deux ensembles de données, comme indiqué dans la section 3.4.1. Ensuite, la variable cible, correspondant à l'état de réservation, est créée dans la section 3.4.2. Par ailleurs, les données sont nettoyées afin d'éliminer toute information susceptible d'affecter la modélisation et de garantir que toutes les variables sont adaptées aux étapes suivantes. Cette étape est détaillée dans la section 3.4.3.

Les deux dernières étapes ont été appliquées à l'ensemble de données D_{2023} qui a été préparé

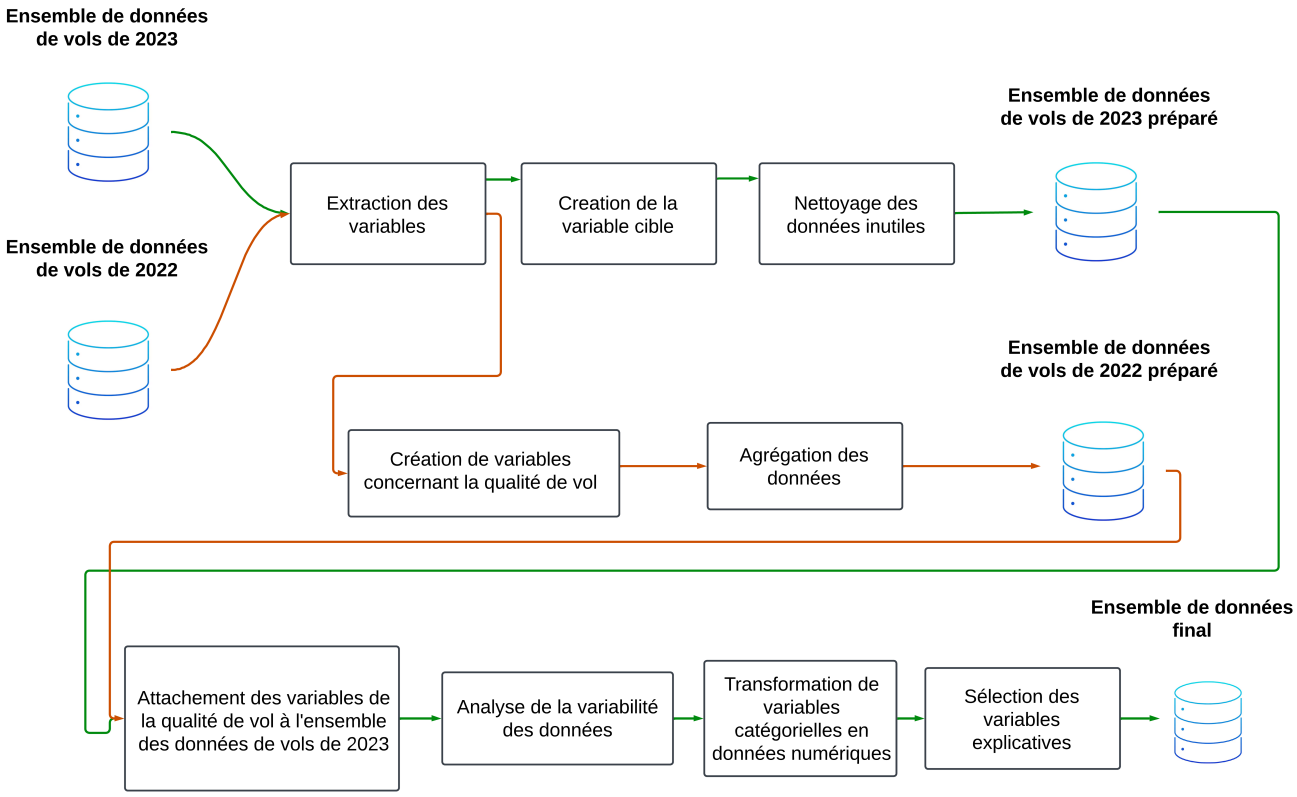


FIGURE 3.1 Méthodologie de préparation des données

à ce niveau. Pour l'ensemble D_{2022}, le processus se poursuit avec la création de variables indiquant la qualité des vols en termes de réservation de surplus, comme décrit dans la section 3.4.4. Ensuite, une agrégation des données est effectuée à un niveau spécifique en préparation de l'étape suivante, comme détaillé dans la section 3.4.5. Une fois les deux ensembles de données préparés, les variables de qualité de vol extraites de D_{2022} sont jointes aux données de vols de 2023, comme expliqué dans la section 3.4.6.

Cet nouvel ensemble contient donc les variables relatives aux réservations de surplus pour les vols de 2023 en combinaison avec les variables historiques, indiquant la qualité des vols en termes de réservation de surplus, pour les vols partant en 2022. En outre, une analyse de la variabilité a été effectuée dans la section 3.4.7. Puis, les variables catégorielles sont transformées en données numériques à l'aide de l'encodage d'étiquettes ou de l'encodage one-hot, ce qui est expliqué dans la section 3.4.8. Ensuite, les variables à utiliser sont sélectionnées sur la base de leur corrélation, comme expliqué dans la section 3.4.9, afin d'obtenir un ensemble de données final D_f prêt pour la modélisation. Enfin, D_f est divisé par région puisque la modélisation est effectuée par région, comme démontré dans la section 3.4.10.

3.2 Analyse exploratoire des données relatives à la problématique

Le programme de fidélité, Aéroplan, reçoit d'Air Canada un nombre spécifique d'allocations de surplus, qui peut augmenter ou diminuer à chaque date de publication jusqu'à la date de départ du vol. Pour chaque vol, il existe 365 dates de publication, où le nombre d'allocations et le nombre cumulatif de réservations de surplus changent. L'objectif de cette analyse exploratoire est de fournir des informations à propos de la difficulté de réserver des sièges de surplus pour les différents marchés. Dans cette analyse, l'ensemble D_{2023}, contenant tous les vols ayant eu lieu entre le 1er janvier 2023 et le 31 décembre 2023, a été pris en compte. Seuls les sièges de la classe économique ont été considérés.

3.2.1 Analyse du trafic pour différentes régions

La première information porte sur les mouvements de vols entre les régions, c'est-à-dire celles présentant le plus grand nombre de vols. La figure 3.2 présente cinq régions, à commencer par la région Domestique, qui affiche le mouvement le plus important avec 301,2 mille vols pour l'année 2023, suivie de la région des États-Unis avec 205,2 mille vols. Dans le deuxième volet, les régions Atlantique, sud et Pacifique sont identifiées, classées respectivement selon le nombre de vols, pour un total cumulé de 62,8 mille vols.

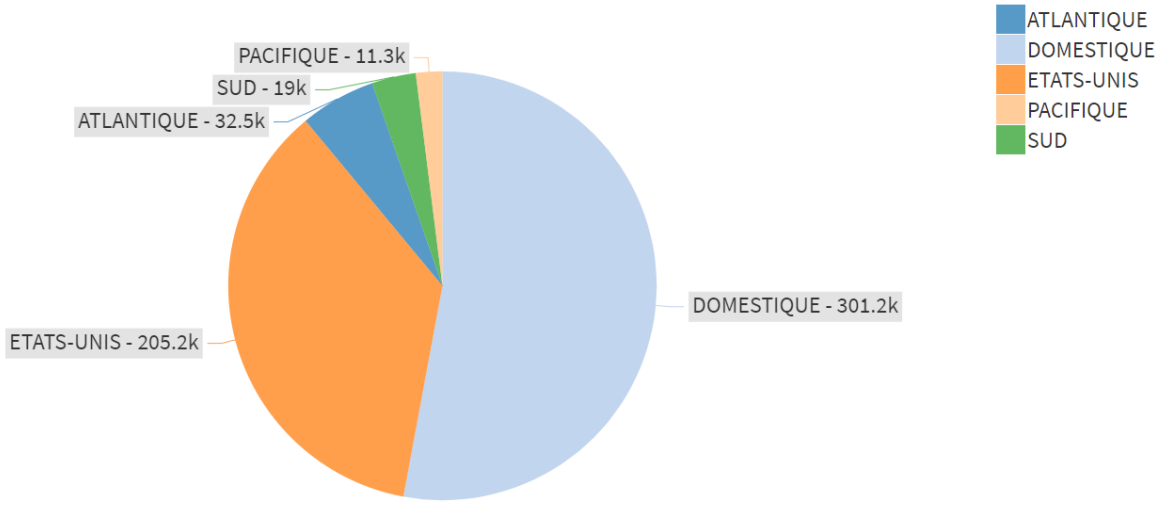


FIGURE 3.2 Nombre de vols par région

Ces vols proviennent de divers itinéraires. La figure 3.3 montre le nombre d'itinéraires associés à chaque région. La région des États-Unis compte un nombre plus élevé d'itinéraires que la région Domestique, avec un total de 243 directions.

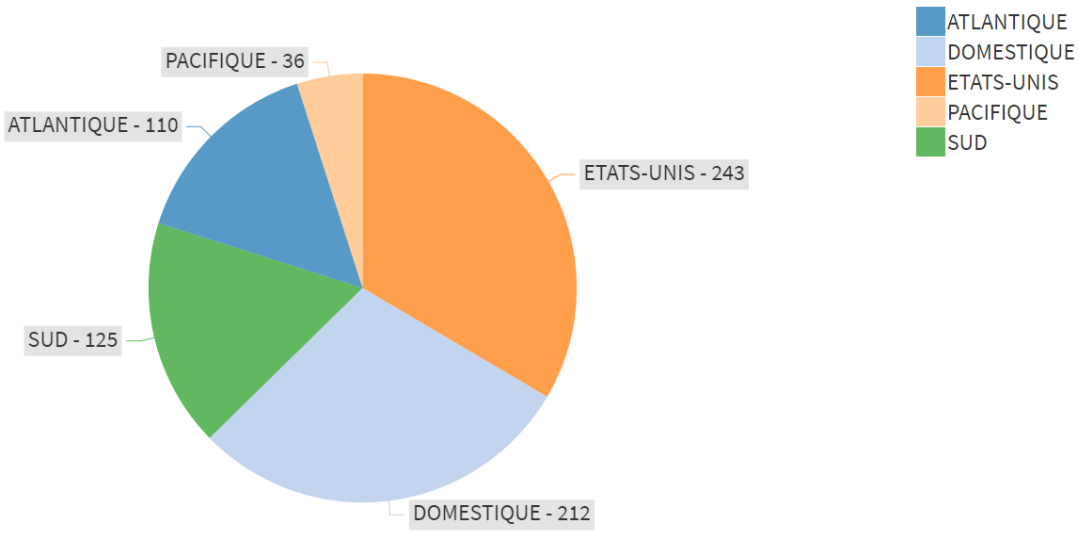


FIGURE 3.3 Nombre d'itinéraires par région

Cela signifie qu'il y a un plus grand nombre de vols par itinéraire dans la région Domestique qu'aux États-Unis. Les régions du Sud et du Pacifique affichent des nombres d'itinéraires proches malgré une différence dans le nombre de vols. La région du Pacifique affiche un total de 36 itinéraires.

3.2.2 Analyse du taux de réservation final et du nombre final d'allocations

Pour illustrer la difficulté de réserver des sièges de surplus, la figure 3.4 montre une analyse du **taux de réservation final (TRF)** moyen par région. Ce taux est calculé comme la moyenne par région du rapport entre le nombre final de réservations et le nombre final d'allocations.

- **Nombre final de réservations** : correspond au nombre total des réservations enregistrées pour un vol à sa date de départ.
- **Nombre final d'allocations (NFA)** : représente le nombre de sièges alloués pour le même vol à sa date de départ .

La figure montre que la région Sud atteint un TRF moyen supérieur à 30%, suivie de la région Pacifique, qui affiche un taux de 26,4%. Les régions États-Unis et Domestique, bien qu'elles comptent un nombre élevé de vols, présentent des taux de réservation finaux moyens plus bas que ceux des régions Sud et Pacifique. La région Atlantique enregistre un taux de 12,1%. Par ailleurs, le TRF moyen pour les sièges de surplus révèle que plus de la moitié des sièges alloués restent non réservés dans toutes les régions.

Ensuite, le pourcentage du nombre total de vols selon les différentes partitions du TRF est examiné. Dans les partitions créées, les valeurs sont égales à zéro dans une seule d'entre elles,

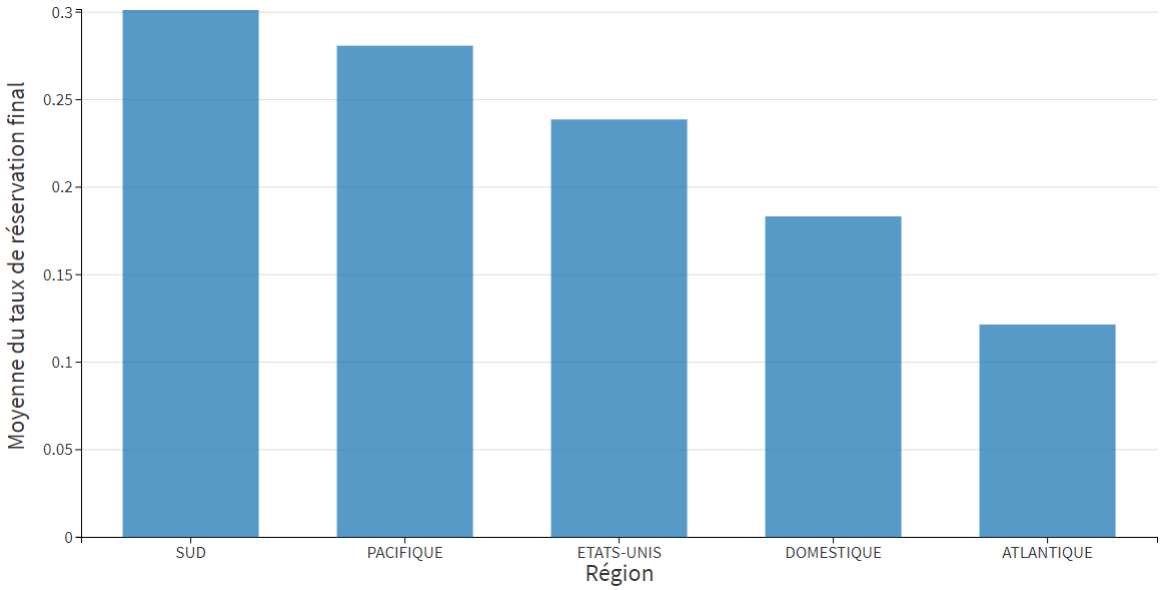


FIGURE 3.4 Moyenne du taux de réservation final par région

car celle-ci présente une majorité de vols comme indiqué dans la figure 3.5. Les partitions sont donc créées de manière à ce que les partitions soient étroits au début et deviennent plus larges au fur et à mesure que le TRF augmente. Une autre partition est crée lorsque le TRF est supérieur à 90%, car il présente une grande pourcentage de vols. Les différentes classes sont définies comme suit :

- **TRF = 0%** : taux de réservation final égal à 0%.
- **0% < TRF ≤ 5%** : taux de réservation final entre 0% et 5%.
- **5% < TRF ≤ 10%** : taux de réservation final entre 5% et 10%.
- **10% < TRF ≤ 20%** : taux de réservation final entre 10% et 20%.
- **20% < TRF ≤ 50%** : taux de réservation final entre 20% et 50%.
- **50% < TRF ≤ 90%** : taux de réservation final entre 50% et 90%.
- **90% < TRF ≤ 100%** : taux de réservation final entre 90% et 100%.

La figure 3.5, montre que 62,1% des vols présentent un TRF égal à 0%, ce qui signifie une difficulté à réserver au moins un siège de surplus pour la plupart des vols. Le deuxième TRF le plus existant, dépassant 90%, représente 18,7% du nombre total de vols. Le reste des vols, soit moins de 20%, présente un TRF entre 0% et 90%, sans inclure les vols dont le TRF est null. Le TRF entre 50% et 90% représente moins de 5% de vols, ce qui confirme la difficulté de réserver plus de la moitié d'allocations de surplus.

La figure 3.6 montre le pourcentage de vols par partition du TRF pour chaque région. Plus de 50% des vols ont un TRF de 0% dans les régions Domestique, Pacifique, Atlantique et des

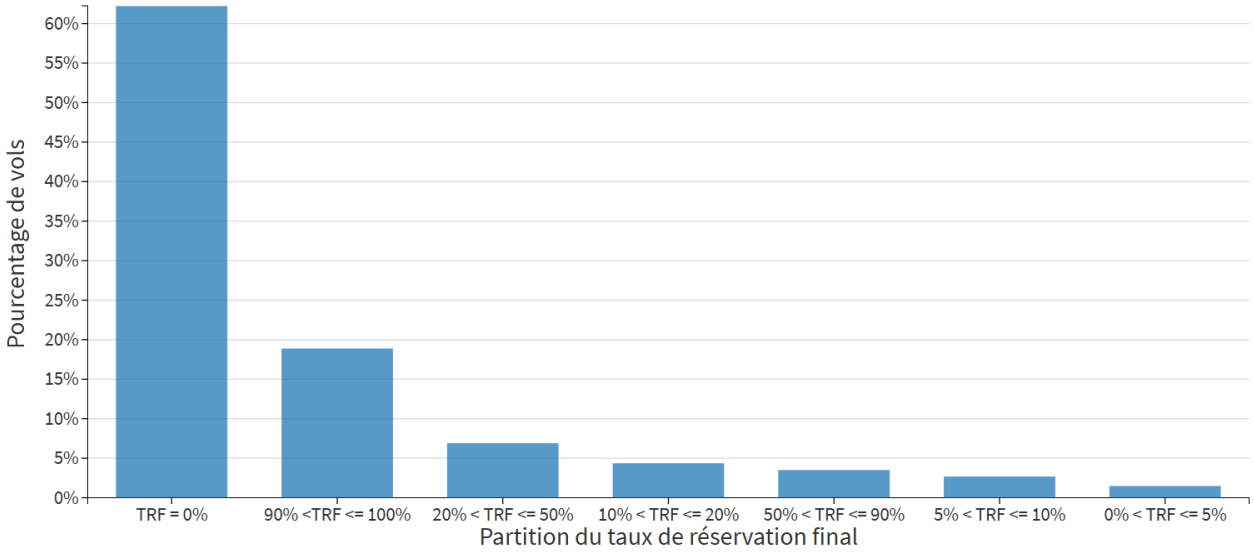


FIGURE 3.5 Pourcentage de vols par partition du taux de réservation final

États-Unis. La région Sud affiche le pourcentage le plus bas, soit 44,2%. Cela indique que le problème de réservation de surplus existe dans toutes les régions. La deuxième partition la plus présente dans toutes les régions est le TRF entre 90% et 100%, ce qui indique plus de 30% des vols dans la région Sud et plus de 20% des vols pour la région Pacifique et les États-Unis.

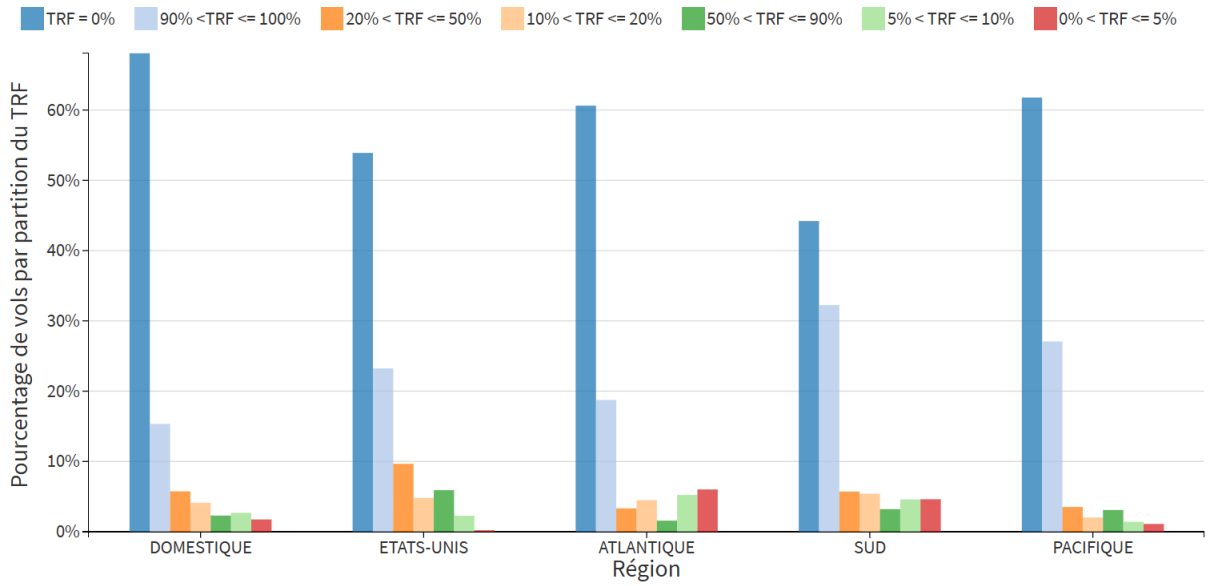


FIGURE 3.6 Pourcentage de vols par partition du taux de réservation final et par région

Dans la figure 3.7, les vols dont le nombre final d'allocations (NFA), tel que défini en 3.2.2, est

nul sont exclus. Cela permet d'éliminer les vols qui n'ont pas reçu d'allocations de surplus, ce qui conduit directement à TRF null permettant de mieux évaluer la demande sur les sièges de surplus. Le pourcentage de vols avec un TRF de 0% diminue dans toutes les régions, en particulier celles du Pacifique et Atlantique. Dans toutes les régions, à l'exception de la région Domestique, le TRF entre 90% et 100% est dominé par un pourcentage de vols dépassant 30%, qui atteint 50% dans la région Pacifique. L'analyse des données montre que l'élimination des vols dont le nombre final d'allocations est nul a un impact sur la répartition du TRF, et qu'un pourcentage élevé de vols dont le TRF est nul est dû au fait qu'ils n'ont pas d'allocations.

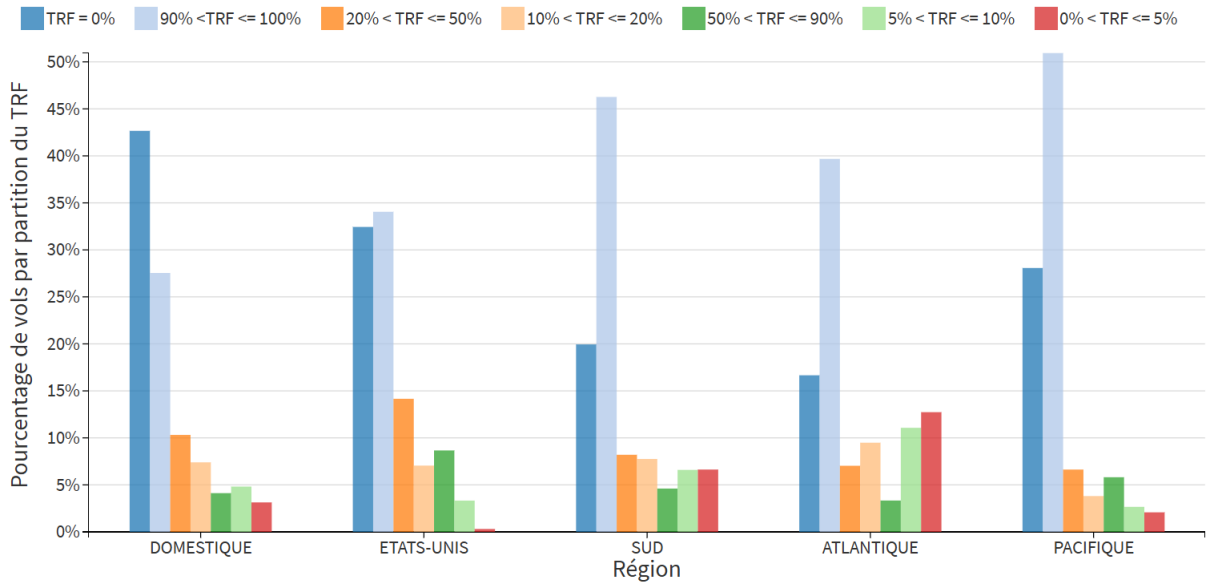


FIGURE 3.7 Pourcentage de vols par partition du taux de réservation final et par région - NFA $\neq 0$

Nous expliquons maintenant la raison derrière le pourcentage élevé de vols présentant un TRF supérieur à 90%. Pour ce faire, le nombre final d'allocations est examiné à l'aide de partitions. Les valeurs de zéro du nombre final d'allocations présentent une seule partition. Les partitions sont plus larges à mesure que le nombre final d'allocations augmente. Les valeurs supérieures à 50 sont placées dans une partition séparée, car elles présentent moins de vols. Les différentes partitions sont définies comme suit :

— **NFA = 0** : nombre final d'allocations égal à 0.
— **0 < NFA ≤ 5** : nombre final d'allocations entre 0 et 5.
— **5 < NFA ≤ 10** : nombre final d'allocations entre 5 et 10.
— **10 < NFA ≤ 25** : nombre final d'allocations entre 10 et 25.
— **25 < NFA ≤ 50** : nombre final d'allocations entre 25 et 50.

— **NFA > 50** : nombre final d'allocations supérieur à 50.

Dans la figure 3.8, la majorité des vols ont un nombre final d'allocations égal à zéro, avec un pourcentage supérieur à 40%. Ceci est expliqué par le fait qu'Air Canada retire les allocations d'Aéroplan pour ces vols ou l'annulation d'un vol. De même, le pourcentage de vols dont le nombre final d'allocations est inférieur à 5 sièges présente 21,3%. Les vols dont le nombre final d'allocations est supérieur à 25 sièges représentent une minorité ne dépassant pas 10% des vols.

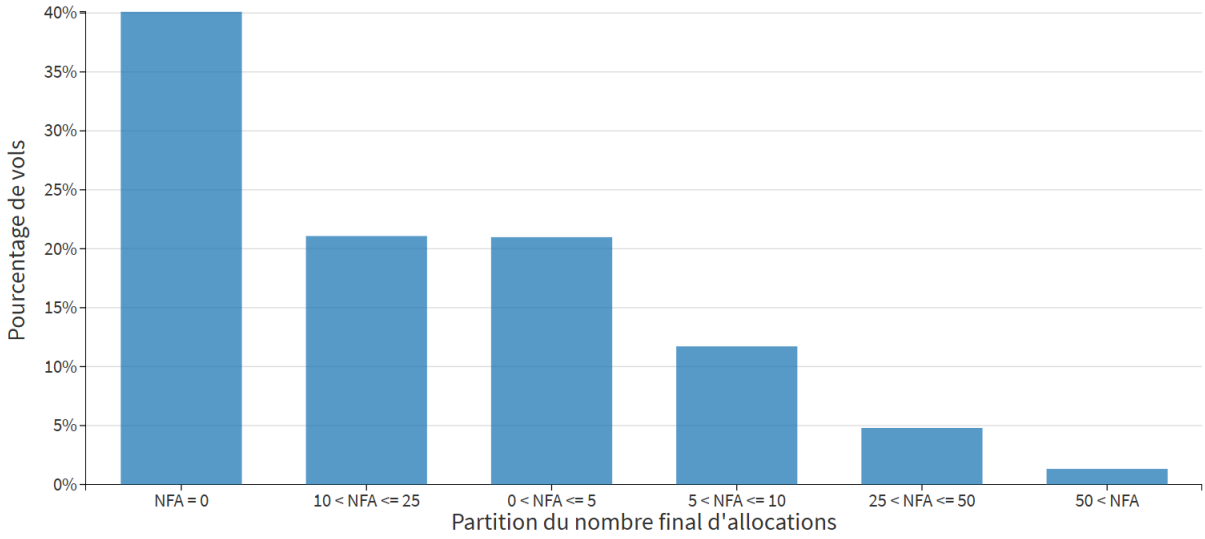


FIGURE 3.8 Pourcentage de vols par partition du nombre final d'allocations

Dans la figure 3.9, tous les vols présentant un nombre final d'allocations égal à zéro sont exclus de l'analyse. Le pourcentage de vols est ensuite observé en fonction des partitions du nombre final d'allocations et des partitions du TRF. Il est constaté que la majorité des vols ayant un TRF compris entre 90% et 100% présentent un faible nombre final d'allocations, souvent inférieur à 5 sièges.

Cela conduit à conclure que la plupart des vols présentant un TRF élevé ont un faible nombre final d'allocations. En revanche, pour un TRF égal à 0, un grand pourcentage de vols avec un nombre final d'allocations compris entre 10 et 25 sièges. Pour un TRF faible, entre 0% et 10%, la proportion la plus élevée de vols présente un nombre final d'allocations supérieur à 10 sièges.

La conclusion est que le TRF seul ne reflète pas la situation des réservations, puisqu'un TRF élevé tend à avoir un NFA plus faible et qu'il est important de vérifier le nombre de réservations. En outre, les vols avec un NFA élevé ont tendance à avoir un TRF faible ou nul, ce qui indique la difficulté de réserver ces sièges.

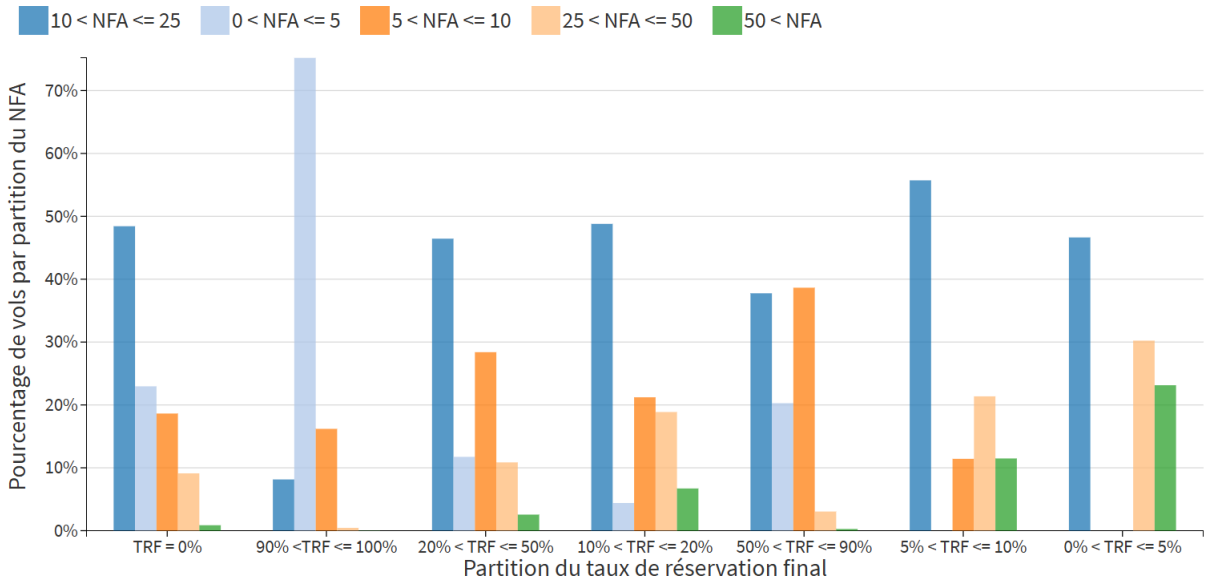


FIGURE 3.9 Pourcentage de vols par partition du taux de réservation final et partition du nombre final d'allocations - NFA $\neq 0$

3.3 Description des données

Dans cette recherche, les données relatives aux réservations de surplus associées aux différents vols sont nécessaires. Ces informations ont été fournies par Air Canada, comme détaillé dans le tableau 3.1. Afin d'améliorer la modélisation de l'état de réservation, deux ensembles de données distincts sont utilisés, chacun répondant à des objectifs différents, expliqués comme suit :

- **Ensemble de données 1** : contient des informations relatives aux réservations de surplus pour les vols au départ de 2023. Cet ensemble constitue la partie principale des données utilisées pour la modélisation de l'état de réservation, y compris l'entraînement et le test des modèles d'apprentissage automatique. Cet ensemble de données est désigné par D_{2023}.
- **Ensemble de données 2** : contient des informations relatives aux réservations de surplus pour les vols au départ de 2022. Cet ensemble est utilisé pour générer de nouvelles variables indiquant la qualité des vols en lien avec la réservation du surplus, comme détaillé dans la section 3.4.4. Cet ensemble de données est désigné par D_{2022}.

Afin d'améliorer la performance de la modélisation de l'état de réservation, D_{2022} est utilisé pour extraire de nouvelles variables historiques relatives à la qualité du vol, qui enrichissent les informations concernant la réservation de surplus, puis les joignons à l'ensemble de données principal D_{2023}. Les nouvelles variables sont extraites de l'ensemble de données D_{2022}, qui

Tableau 3.1 Description des variables extraites

Variable	Étiquette	Définition	Type
Date de départ	Dept_date	Indique le jour, le mois et l'année de départ d'un vol	Date
Itinéraire	OD	Paire origine/destination du vol	Catégorique
Origine	Orig	Ville de départ	Catégorique
Destination	Dest	Ville d'arrivée	Catégorique
Numéro de vol	Flight_num	Numéro qui correspond à un temps exact de départ d'un vol en heures, minutes et secondes	Numérique
Date de publication	Post_date	Indique un jour exact de l'année présentant des informations sur la réservation de sièges et le nombre d'allocations par vol	Date
Fenêtre de réservation	Days_Prior_out	Nombre de jours entre la date de publication et la date de départ	Numérique
Nombre d'allocations	Alloc_for_X	Nombre de sièges de surplus alloués à la réservation à une date de publication	Numérique
Nombre de réservations	Bkg_for_X	Nombre cumulatif de réservations de surplus à une date de publication	Numérique
Disponibilité des sièges	Avail_for_X	Le nombre des sièges de surplus non encore réservées à une date de publication	Numérique
Taux de réservation	BLF_X	Rapport entre le nombre de réservations et le nombre d'allocations à une date de publication	Numérique
Nombre de réservations incrémentale	Book_num_X	Nombre de nouvelles réservations de surplus à une date de publication	Numérique
Jour de la semaine de départ	DOW_departure	-	Catégorique
Jour du mois de départ	day_of_month_dep	-	Numérique
Mois de départ	Departure_month	-	Catégorique
Jour de la semaine de réservation	DOW_booking	Jour de la semaine de la date de publication	Catégorique
Jour du mois de réservation	day_of_month_book	Jour du mois de la date de publication	Numérique
Mois de réservation	Booking_month	Mois de de la date de publication	Catégorique
Heure du temps de départ	Hour_dep	-	Numérique
Minute du temps de départ	Minute_dep	-	Numérique
Saison de départ	Season	Saison à laquelle appartient la date de départ (hiver, été, automne, printemps)	Catégorique
Partie du jour de départ	PartOfDay	Partie de la journée à laquelle appartient l'heure de départ	Catégorique
Distance du vol	Flight_distance	Distance parcourue entre origine et destination	Numérique
Région	Region_code	La région à laquelle appartient l'itinéraire	Catégorique

contient d'anciennes informations de réservation de surplus sur les vols en 2022, afin d'éviter toute distorsion ou biais dans le processus de modélisation.

La modélisation de l'état de réservation du surplus se fait par région, car elles représentent des marchés différents pour la compagnie aérienne. Les différentes régions existantes sont les suivantes : Domestique, États-Unis (USA), Atlantique, Sud et Pacifique. Ces marchés desservent un large éventail d'itinéraires. Chaque itinéraire présente des caractéristiques spécifiques concernant les réservations de vols. Une attention particulière est portée à la classe économique en raison de son taux de réservation élevé par rapport à la classe économique premium et à la classe affaires.

3.4 Application du processus de préparation des données

Cette section présente les étapes qui ont été appliquées aux ensembles de données D_{2023} et D_{2022}. Toutes les étapes ont été effectuées, telles que le nettoyage des informations inutiles et biaisées, la création de nouvelles variables, ainsi que la transformation des variables catégorielles. L'extraction des variables constitue la première étape du processus.

3.4.1 Extraction des variables

Dans cette étape, les variables présentes dans les deux ensembles de données, telles que celles répertoriées dans le tableau 3.1, sont identifiées. Ensuite, les données utiles à la recherche sont examinées, notamment les informations sur la date de départ, la fenêtre de réservation, la date de publication, l'origine, la destination, ainsi que la variation du nombre d'allocations et de réservations de surplus à chaque date de publication. La même collecte de variables est effectuée dans les deux ensembles de données.

3.4.2 Création de la variable cible

Pour l'ensemble de données D_{2023}, la variable cible y, représentant l'état de réservation, a été construite. Cette variable dépendante, notée `STATE_OF_BOOKING`, indique si au moins une réservation de surplus a été effectuée dans les sept jours suivants. Pour chaque vol, cette variable est calculée à chaque date de publication. Il s'agit d'une variable binaire prenant la valeur 0 en l'absence de nouvelle réservation dans les sept jours suivants, et la valeur 1 lorsqu'au moins une nouvelle réservation est enregistrée dans ce même intervalle de temps.

Nous supposons que pour un vol spécifique, x_r est la variable du nombre de réservations incrémentale, définie dans le tableau 3.1, et e est un indice d'échelonnement de fenêtre de

réservation allant en ordre décroissant de 364 à 0 jours, où 0 indique la date de départ et 364 est la première date de publication. La variable cible y_e à e jours avant la date de départ est définie comme suit :

$$y_e = \begin{cases} 1 & \text{si } \sum_{o=0}^{6} x_{r,e-o} > 0, \\ 0 & \text{sinon.} \end{cases} \tag{3.1}$$

Deux classes sont alors définies :

- **Classe 0 (négative)** : Pas de réservation dans les sept prochains jours (classe majoritaire).
- **Classe 1 (positive)** : Réservation d'au moins un siège dans les sept prochains jours (classe minoritaire).

3.4.3 Nettoyage des informations inutiles

La première étape consiste à supprimer les données dupliquées dans l'ensemble de données et à exclure les cas où aucun siège n'est disponible pour la réservation. Il s'agit notamment des situations où toutes les allocations sont déjà réservées ou lorsqu'il n'y a pas d'allocations du tout. Ces données ont une influence sur les résultats, car l'absence de sièges disponibles entraîne automatiquement un état de réservation de classe majoritaire, c'est-à-dire forte corrélation, sans tenir compte des autres variables. Ce processus permet non seulement d'améliorer la structure des données, mais aussi de réduire le nombre d'échantillons indiquant un état de réservation de classe majoritaire, ce qui signifie que le rapport entre le nombre d'échantillons de la classe minoritaire et celui de la classe majoritaire est augmenté.

Le tableau 3.2 présente le nombre total d'échantillons et le pourcentage de la classe minoritaire dans cinq régions avant et après le nettoyage des informations inutiles. Le pourcentage de la classe minoritaire est la proportion d'échantillons appartenant à la classe 1 par rapport au nombre total d'échantillons. Cette analyse est relative au D_{2023}, qui présente le principal ensemble de données. Les résultats montrent qu'il existe des données déséquilibrées pour la variable cible, ce qui signifie que la distribution des classes dans la variable cible est asymétrique avec un faible pourcentage de la classe minoritaire. Initialement, le nombre d'échantillons était élevé, notamment pour les régions Domestique par 73,9 millions et États-Unis par 51,8 millions, avec des pourcentages de la classe minoritaire faibles variant entre 2% et 3%. Après la préparation des données, le nombre d'échantillons a diminué, en particulier pour les régions Domestique et États-Unis, tandis que les pourcentages de la classe minoritaire ont augmenté. Des améliorations ont été observées dans la région Atlantique, où la pourcentage de la classe minoritaire est passé de 3% à 23%, et dans la région Pacifique, qui a vu une augmentation de 3% à 17%. Globalement, la préparation des données a permis

d'obtenir des ensembles de données plus ciblés et moins déséquilibrés dans toutes les régions. Après cette étape, l'ensemble D_{2023} est bien préparé.

TABLEAU 3.2 Informations sur les pourcentages de la classe minoritaire par région pour les vols de 2023

Région	Nombre d'itinéraires	Avant nettoyage des données		Après nettoyage des données	
		Nombre d'échantillons	% de la classe minoritaire	Nombre d'échantillons	% de la classe minoritaire
Domestique	212	73 909 079	2%	26 279 229	7%
États-Unis	243	51 805 146	2%	16 865 371	8%
Sud	125	5 071 404	3%	1 224 857	16%
Pacifique	36	2 875 794	3%	618 829	17%
Atlantique	110	9 867 054	3%	1 014 639	23%

3.4.4 Création de variables relatives à la qualité du vol

L'ensemble D_{2023} contient des informations au niveau de la date de publication, indiquant la variation du nombre d'allocations, le nombre cumulatif de réservations, le taux de réservation et la fenêtre de réservation. Ces données fournissent des informations relatives à la réservation de surplus. Au niveau du vol, D_{2023} inclut des variables telles que l'itinéraire, la région, la date de départ et la distance parcourue par le vol qui sont des informations qui ne caractérisent que le vol sans aucune information sur la réservation de surplus.

L'objectif de cette étape est d'enrichir D_{2023} avec de nouvelles données liées aux réservations de surplus au niveau du vol présentées dans le tableau 3.3, appelées variables de la qualité du vol. Parmi ces nouvelles variables figurent :

- Nombre maximal et final de réservations par vol.
- Nombre maximal et final d'allocations par vol.
- Taux de réservation final et maximal par vol.
- Nombre de jours présentant de nouvelles réservations par vol.
- Nombre de jours présentant des allocations par vol.

Ces variables permettent d'identifier les vols qui ont tendance à réserver davantage de surplus ou à bénéficier de plus d'allocations de surplus que d'autres. Cela les désigne comme des vols prioritaires à gérer.

Afin d'éviter tout biais dans la phase de modélisation, les informations relatives à la qualité des vols n'ont pas été extraites de l'ensemble principal D_{2023}, utilisé pour l'entraînement du modèle. Lorsque le modèle prédictif a été construit, il est fondamental que les variables explicatives utilisées ne soient pas influencées par des informations futures ou directement dérivées de la variable cible à prédire. Si l'on utilisait, pour enrichir D_{2023}, des variables résumant des comportements de réservation sur la même période (2023), cela reviendrait à

TABLEAU 3.3 variables de la qualité du vol

Variable	Étiquette	Définition	Type
Taux de réservation maximale par vol	Max_blf	Le plus grand taux de réservation	Numérique
Nombre maximale d'allocations par vol	Max_alloc	Le plus grand nombre d'allocations	Numérique
Nombre maximale de réservations par vol	Max_book	Le plus grand nombre cumulatif de réservations	Numérique
Taux de réservation final par vol	Final_blf	Taux de réservation à la date de départ	Numérique
Nombre final d'allocations par vol	Final_alloc	Nombre d'allocations à la date de départ	Numérique
Nombre final de réservations par vol	Final_book	Nombre de réservations à la date de départ	Numérique
Nombre de jours sans réservation par vol	Days_without_book	Nombre de jours où aucune nouvelle réservation n'a été effectuée	Numérique
Nombre de jours présentant de réservation par vol	Days_with_book	Nombre de jours avec au moins une nouvelle réservation	Numérique
Nombre de jours présentant de disponibilité par vol	Num_days_with_Avail	Nombre de jours où il reste au moins un siège à réserver	Numérique
Nombre de jours présentant d'allocations par vol	Num_days_with_alloc	Nombre de jours avec au moins un siège alloué	Numérique

injecter des informations connues à l'avance, ce qui biaiserait la modélisation et donnerait une illusion de performance : on parle alors de fuite de données.

À la place, un ensemble distinct, D_{2022}, correspondant à une autre année de départ des vols, a été défini et exploité. Celui-ci a servi exclusivement à identifier des tendances générales sur les comportements de réservation de surplus, sans introduire d'informations postérieures dans le processus de modélisation. Cette approche permet à la fois d'enrichir les données de modélisation avec des variables pertinentes, tout en préservant l'intégrité et la robustesse du modèle prédictif.

De plus, une autre variable, appelée vitesse de réservation, est créée. Celle-ci est définie comme suit :

— **Vitesse de réservation** : est le nombre de jours passés depuis la dernière réservation. Les jours où il n'y a pas de disponibilités d'allocations ne sont pas comptés. Plus la valeur de cette variable est élevée, plus le temps nécessaire pour réserver est long, et inversement.

Pour mieux illustrer la façon dont cette variable varie selon la fenêtre de réservation, le

tableau 3.4 fournit un exemple d'une partie des données qui représente ce variation.

Par exemple, pour une fenêtre de réservation de 222 jours, la vitesse de réservation est de 7, ce qui indique que la réservation la plus récente a été effectuée 7 jours plus tôt, à une fenêtre de réservation de 229 jours. L'état de la réservation est marqué comme étant classe minoritaire car il y a eu 4 réservations de surplus enregistrées à une fenêtre de réservation de 216 jours.

En outre, pour une fenêtre de réservation comprise entre 216 et 222 jours, la variable cible indique la classe minoritaire, étant donné qu'une nouvelle réservation de surplus a lieu dans les sept jours suivants, en particulier à une fenêtre de réservation de 216 jours.

Tableau 3.4 Variation de la vitesse de réservation selon la fenêtre de réservation

Fenêtre de réservation	Nombre d'allocations	Réservations incrémentales	Vitesse de réservation	État de réservation
230	10	0	-	1
229	10	2	-	1
228	10	0	1	0
227	10	0	2	0
226	10	0	3	0
225	10	0	4	0
224	12	0	5	0
223	12	0	6	0
222	12	0	7	1
221	12	0	8	1
220	16	0	9	1
219	16	0	10	1
218	16	0	11	1
217	19	0	12	1
216	19	4	-	1
215	19	0	1	0
214	21	0	2	0

En raison d'une forte corrélation observée entre la vitesse de réservation et la variable cible, une transformation de cette variable a été nécessaire afin de préserver son pouvoir explicatif tout en réduisant sa dépendance directe au résultat à prédire. Pour ce faire, une discrétisation de la vitesse de réservation a été effectuée en fonction des partitions de la fenêtre de réservation.

Cette transformation repose sur une segmentation prédéfinie des fenêtres de réservation, élaborée par Air Canada, qui reflète les comportements typiques de réservation de surplus : plus la date de départ approche, plus les réservations de surplus sont fréquentes. Ainsi, les intervalles deviennent plus étroits à l'approche de la date de départ, comme le montre le tableau 3.5.

La vitesse de réservation moyenne, notée `Speed_Partition`, a ensuite été calculée pour chaque classe, permettant ainsi de créer une variable discrète qui capture le rythme des ré-

Tableau 3.5 Répartition des fenêtres de réservation

Partition	Fin de la fenêtre	Début de la fenêtre
1	355	332
2	331	278
3	277	244
4	243	215
5	214	187
6	186	158
7	157	124
8	123	96
9	95	68
10	67	47
11	46	36
12	35	24
13	23	16
14	15	9
15	8	6
16	5	3
17	2	2
18	1	1
19	0	0

servations au sein de chaque intervalle de temps. Cette nouvelle variable permet de conserver l'information liée à la dynamique de réservation tout en limitant les risques de surapprentissage liés à une forte corrélation avec la variable cible.

3.4.5 Agrégation des variables de la qualité de vol

Cette étape vise à préparer la jointure des variables de qualité de vol extraites de l'ensemble D_{2022} vers les vols de l'année 2023 dans D_{2023}. Il est observé que les numéros de vol, qui sont des désignations de l'heure exacte de départ, diffèrent généralement d'une année à l'autre, même lorsque l'itinéraire et la date de départ sont similaires. Ainsi, le numéro de vol ne constitue pas une clé fiable pour identifier des correspondances entre les deux années.

Afin d'établir une correspondance robuste entre les vols de D_{2022} et ceux de D_{2023}, la jointure a été réalisée à l'aide de variables plus stables : l'itinéraire, la date de départ (jour et mois), la partie du jour de départ (segment horaire de départ) et la partition du nombre final d'allocations.

Le choix de la partie du jour à la place de l'heure exacte de départ repose sur l'objectif de

réduire la granularité temporelle et de limiter les erreurs de correspondance dues à des variations d'horaire. En effet, deux vols ayant lieu dans des plages horaires similaires (ex. : matin ou après-midi) sont considérés comme comparables en termes de comportement opérationnel et de qualité. Cela permet de regrouper les vols dans des intervalles temporels homogènes. L'ajout de la partition du nombre final d'allocations vise à affiner cette correspondance en tenant compte de la capacité ou du niveau de l'allocation des sièges de surplus. Deux vols proches dans le temps peuvent en effet présenter des caractéristiques différentes si leur niveau d'allocations varie fortement. En intégrant cette partition, la comparaison devient plus cohérente et spécifique aux contextes opérationnels similaires.

Cependant, plusieurs vols de D_{2022} peuvent correspondre à un vol donné de D_{2023} selon ces critères. Dans ce cas, les variables de qualité sont agrégées par groupe, à l'aide de leur médiane, calculée comme suit :

$$x_{\text{agg}} = \text{médiane}\left(x_1, x_2, \ldots, x_{n_{\text{agg}}}\right), \tag{3.2}$$

où n_{agg} désigne le nombre de vols correspondant.

3.4.6 Jointure des variables de qualité de vol à l'ensemble D_{2023}

Dans cette section, les informations historiques relatives à la qualité des vols en termes de réservation de surplus sont jointes à l'ensemble des données D_{2023}. Cette jointure est effectuée au niveau des variables suivantes : itinéraire, jour et mois de départ, partition du nombre final d'allocations, et partie du jour de départ. Un ensemble de données est ainsi obtenu, contenant à la fois les informations sur les vols actuels en 2023 et les informations historiques des vols de 2022 concernant la réservation de surplus, au niveau du vol ou même au niveau de la fenêtre de réservation grâce à la partition de la vitesse de réservation. La figure 3.10 illustre la jointure réalisée dans D_{2023}.

3.4.7 Analyse de la variabilité des données

Dans cette section, la variabilité des données dans l'ensemble de données D_{2023}, après la jointure des variables de qualité de vol, est interprétée dans le tableau 3.6. Cet ensemble contient $n = 46\,002\,925$ échantillons. La moyenne, l'écart-type et l'asymétrie de chaque variable numérique sont observés, et sont calculés respectivement comme suit :

$$\mu_{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \tag{3.3}$$

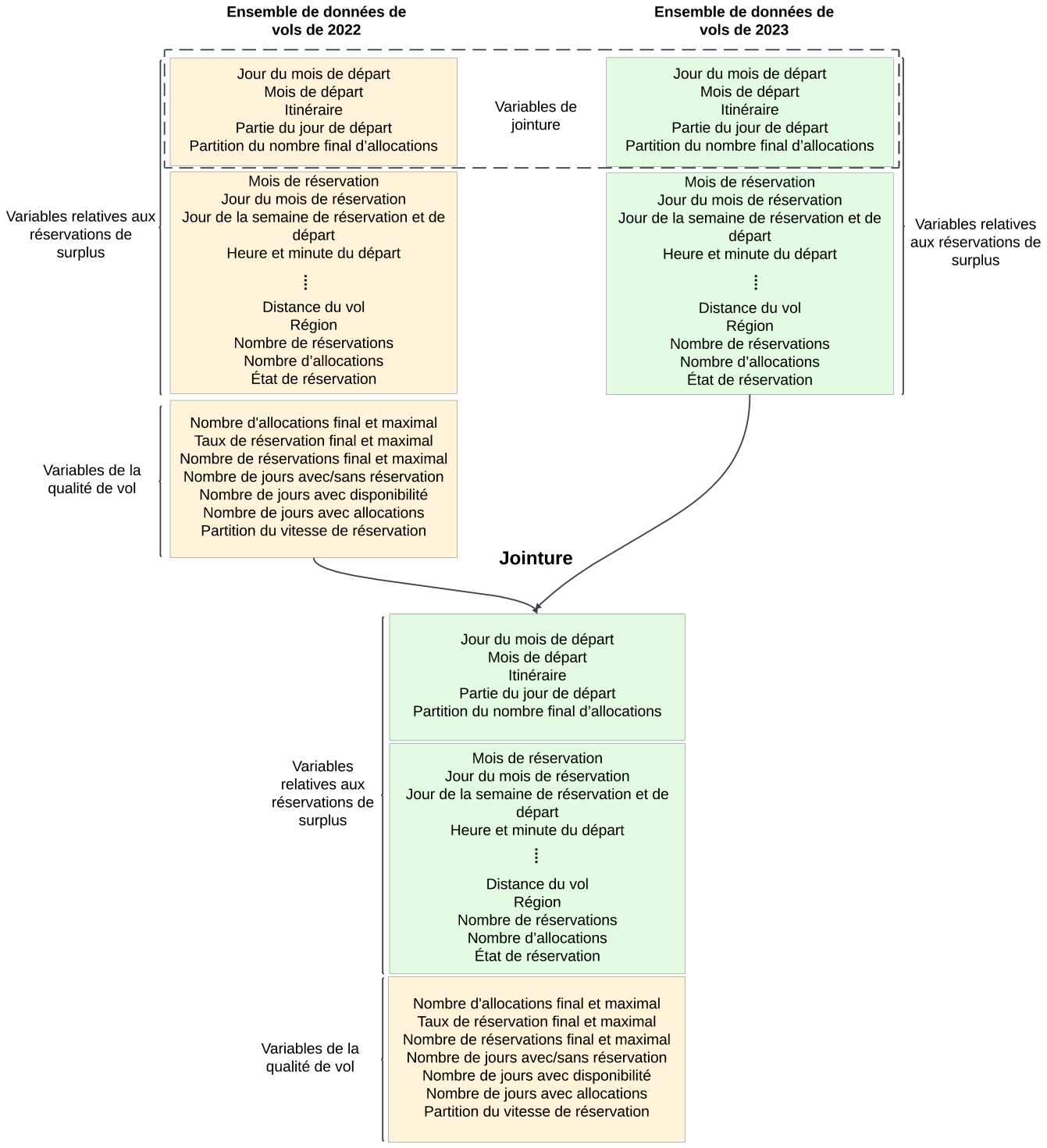


FIGURE 3.10 Jointure des variables de qualité des vols à D_{2023}

$$\sigma_{x_j} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \mu_{x_j})^2}. \tag{3.4}$$

$$\text{Asymétrie}(x_j) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{ij} - \mu_{x_j}}{\sigma_{x_j}}\right)^3 \tag{3.5}$$

L'asymétrie mesure la symétrie d'une distribution par rapport à sa moyenne. Une valeur d'asymétrie égale à 0 indique une distribution parfaitement symétrique, comme la distribution normale. Une asymétrie positive signifie que la distribution est asymétrique à droite : la majorité des valeurs sont concentrées vers la gauche avec une longue queue vers la droite. À l'inverse, une asymétrie négative indique une asymétrie à gauche, où les valeurs sont regroupées vers la droite avec une longue queue vers la gauche.

Tableau 3.6 Analyse descriptive et asymétrie des variables numériques

Variable	Moyenne	Ecart-type	Asymétrie	Interprétation
DAYS_PRIOR_OUT	227.88	96.08	-0.67	Distribution modérément asymétrique à gauche, avec majorité des valeurs proches de 200 jours avant le départ.
ALLOC_FOR_X	18.57	14.29	1.30	Distribution asymétrique à droite, certains vols ayant des allocations plus élevées.
AVAIL_FOR_X	18.11	14.10	1.24	Similaire à ALLOC_FOR_X
BKG_FOR_X	0.46	1.63	6.92	Extrêmement asymétrique à droite, la plupart des valeurs étant proches de 0.
DOW_BOOKING	3.00	2.00	0.00	Répartition uniforme sur les jours de semaine.
DOW_DEPARTURE	3.11	1.92	-0.09	Distribution quasi symétrique, départs également répartis.
BOOKING_MONTH	6.31	3.41	0.06	Uniformité dans la répartition des mois de réservation.
DEPARTURE_MONTH	5.89	3.58	0.16	Légère asymétrie à droite, mais proche d'une distribution uniforme.
DAY_OF_MONTH_DEP	15.54	8.75	0.02	Répartition homogène dans le mois.
DAY_OF_MONTH_BOOK	15.77	8.76	0.00	Répartition homogène.
HOUR_DEP	12.45	4.98	0.21	La moyenne proche de midi indique des départs fréquents en milieu de journée
MINUTE_DEP	18.50	17.92	0.68	Légère asymétrie à droite, horaires un peu concentrés.
FLIGHT_DISTANCE	466.22	600.87	4.83	Très forte asymétrie à droite, la plupart des vols sont courts mais quelques-uns très longs.
STATE_OF_BOOKING	0.08	0.27	3.11	La plupart des cas sont à 0 (pas de réservéation). **Variable très déséquilibrée**.
SPEED_PARTITION	345.32	58.41	-3.88	Asymétrique à gauche, longues périodes depuis la dernière réservation.
DAYS_WITHOUT_BOOK	306.46	64.44	-1.38	Longues périodes sans réservation, à tendance asymétrique à gauche.
DAYS_WITH_BOOK	5.29	6.33	1.94	Majorité de cas avec très peu de jours avec réservations.
NUM_DAYS_WITH_ALLOC	263.29	89.10	-0.73	Distribution légèrement asymétrique à gauche. Les périodes avec allocations sont longues.
NUM_DAYS_WITH_AVAIL	249.89	88.13	-0.63	Distribution légèrement asymétrique à gauche. Les périodes avec disponibilité d'allocations sont longues
MAX_BOOK	3.15	4.38	2.98	Grand nombre de réservations observé seulement dans quelques cas.
MAX_ALLOC	37.86	31.10	2.15	Distribution asymétrique à droite, certains vols avec grand nombre d'allocations.
MAX_BLF	0.44	0.58	4.59	Asymétrique, certains vols avec taux de réservation maximale proches de 1.
FINAL_BLF	0.26	0.40	1.83	Moyenne faible avec dispersion, asymétrie à droite.
FINAL_ALLOC	17.02	16.28	1.39	Distribution modérément asymétrique à droite.
FINAL_BOOK	2.59	4.03	3.18	Faible volume de réservations finales, avec certains cas atteignant des pics.

Afin de réduire l'asymétrie marquée de certaines variables, une transformation logarithmique

a été appliquée à l'aide de la fonction `log1p`, telle que présentée dans l'équation 3.6.

$$\text{log1p}(x_j) = \log(1 + x_j). \tag{3.6}$$

La réduction de l'asymétrie des variables permet d'améliorer la qualité des séparations effectuées par les arbres dans LightGBM. Une variable asymétrique peut concentrer ses valeurs dans un intervalle restreint, ce qui limite les seuils de classification efficaces. Cela peut aussi provoquer un surapprentissage sur des valeurs extrêmes et influencer les mesures d'importance des variables. En appliquant la transformation `log1p`, la distribution devient plus étalée, ce qui facilite la construction des arbres et améliore la stabilité du modèle.

Cette transformation permet de stabiliser la variance et de rendre la distribution plus symétrique. Elle a été appliquée aux variables présentant une forte asymétrie telles que le nombre final et maximal de réservations, le nombre de jours avec réservations, le nombre d'allocations final et maximal, la vitesse de réservation moyenne, le nombre d'allocations ainsi que la distance du vol. Les résultats montrent une réduction de l'asymétrie dans le tableau 3.7.

Tableau 3.7 Comparaison des statistiques avant et après transformation `log1p`

Variable	Avant transformation			Après transformation		
	Moyenne	Écart-type	Asymétrie	Moyenne	Écart-type	Asymétrie
ALLOC_FOR_X	18.57	14.29	1.3	2.69	0.78	-0.2
SPEED_PARTITION	345.32	58.41	-3.88	5.48	0.5	-0.46
FINAL_BOOK	2.59	4.03	3.18	0.83	0.89	0.6
MAX_BOOK	3.15	4.38	2.98	0.99	0.91	0.35
FLIGHT_DISTANCE	466.22	600.87	4.83	5.77	0.75	0.97
FINAL_ALLOC	17.02	16.28	1.39	2.32	1.23	-0.62
MAX_ALLOC	37.86	31.10	2.15	3.43	0.63	0.52
AVAIL_FOR_X	18.11	14.10	1.24	2.65	0.8	-0.24
BKG_FOR_X	0.46	1.63	6.92	0.18	0.48	2.81
DAYS_WITHOUT_BOOK	306.46	64.44	-1.38	5.69	0.27	-0.71
DAYS_WITH_BOOK	5.29	6.33	1.94	1.33	1.04	0.03

3.4.8 Transformation des variables catégorielles en données numériques

Dans cette partie, deux techniques d'encodage de variables catégorielles sont mises en œuvre pour préparer les données en vue de leur utilisation dans des modèles d'apprentissage automatique : l'encodage des étiquettes et l'encodage *one-hot*. Ces techniques permettent de convertir des données catégorielles en valeurs numériques tout en tenant compte des particularités de chaque variable. L'encodage des étiquettes est utilisé pour transformer des catégories ordon-

nées en valeurs discrètes, tandis que l'encodage *one-hot* est appliqué aux variables sans ordre inhérent, évitant ainsi d'introduire des relations fausses entre les catégories. Ces approches garantissent une représentation adéquate des données pour les algorithmes de modélisation.

Encodage des étiquettes

L'encodage des étiquettes a été utilisé pour transformer une variable catégorielle en une variable numérique. Il s'agit de variables telles que le jour de la semaine de départ et de réservation, le mois de départ et de réservation et la partie du jour de départ. Soit une variable catégorielle x_j qui prend des valeurs dans l'ensemble $\{g_1, g_2, \ldots, g_U\}$, où U est le nombre total de catégories distinctes, et chaque g_h représente une catégorie.

L'objectif de l'encodage des étiquettes est de mapper chaque catégorie g_h à une étiquette numérique l_h unique. Ce processus peut être formalisé par une fonction F qui associe chaque catégorie g_h à une étiquette l_h, telle que :

$$F(g_h) = l_h, \tag{3.7}$$

où :

$$l_h \in \{0, 1, 2, \ldots, U-1\}$$

Ainsi, les catégories $\{g_1, g_2, \ldots, g_U\}$ sont remplacées par les étiquettes numériques $\{0, 1, 2, \ldots, U-1\}$. Dans la figure 3.11, un exemple des variables catégorielles avec leurs valeurs est présenté, puis le résultat de leur encodage en variables numériques en transformant leurs étiquettes en nombres discrets à partir de zéro et en prenant un ordre croissant basé sur le temps, soit dans l'ordre des jours, des mois ou d'une partie de la journée.

Soit $x_j = \{x_{1,j}, x_{2,j}, \ldots, x_{n,j}\}$ l'ensemble des valeurs prises par la variable catégorielle x_j dans les n échantillons, où chaque $x_{i,j}$ représente la valeur observée pour l'échantillon i, avec $i \in \{1, 2, \ldots, n\}$. Après l'application de la fonction d'encodage F, chaque valeur catégorielle $x_{i,j}$ est transformée en une étiquette $l_{i,j}$. La variable catégorielle devient :

$$x_j = \{l_{1,j}, l_{2,j}, \ldots, l_{n,j}\}.$$

où $l_{i,j} = F(x_{i,j})$, et chaque $l_{i,j}$ est une valeur numérique correspondant à la valeur catégorielle originale $x_{i,j}$.

L'encodage des étiquettes est caractérisé par :

— **Unicité du mappage** : chaque catégorie g_h est associée à une étiquette numérique

Jour de la semaine de départ

g_h	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
l_h	0	1	2	3	4	6	7

Jour de la semaine de réservation

g_h	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
l_h	0	1	2	3	4	6	7

Partie du jour de départ

g_h	Morning	Afternoon	Evening	Night
l_h	0	1	2	3

Mois de la réservation ou du départ

g_h	January	February	March	April	May	June	July	August	September	October	November	December
l_h	0	1	2	3	4	5	6	7	8	9	10	11

FIGURE 3.11 Encodage des variables catégorielles en variables numériques

unique l_h, garantissant qu'aucune étiquette n'est partagée par deux catégories distinctes.

— **Plage des étiquettes** : les étiquettes se situent dans l'intervalle $\{0, 1, \ldots, U-1\}$, où U est le nombre total de catégories uniques.

— **Ordre implicite** : les étiquettes sont des entiers, ce qui introduit implicitement un ordre entre les catégories. Cependant, cet ordre est artificiel et ne reflète pas nécessairement un ordre naturel dans les données catégorielles d'origine.

Encodage one-hot

La deuxième technique appliquée est l'encodage *one-hot* qui est une technique utilisée pour représenter des variables catégorielles sous forme de vecteurs binaires. Il est utilisé en apprentissage automatique lorsque les algorithmes nécessitent des entrées numériques, et il permet d'éviter les problèmes causés par l'encodage des étiquettes, notamment lorsque les catégories n'ont pas d'ordre inhérent.

Les variables d'origine, destination et de région présentent des valeurs non ordonnées. Par conséquent, elles ne peuvent pas être traitées par l'encodage des étiquettes, car cela les transformerait en valeurs incrémentielles, créant ainsi des importances différentes pour ces valeurs et biaisant le modèle en conséquence. La solution appropriée consiste donc à appliquer l'encodage *one-hot*, car il n'introduit pas de fausses relations ordinales entre les catégories.

L'encodage *one-hot* transforme une variable catégorielle avec U catégories uniques en U variables binaires. Chaque catégorie est représentée par un vecteur binaire où un seul élément est égal à 1 (*hot*) et tous les autres éléments sont égaux à 0.

Le processus de l'encodage one-hot est comme suit :

1. **Identification de la variable catégorielle** : supposons que la variable catégorielle x_j contienne des valeurs provenant de l'ensemble $\{g_1, g_2, \ldots, g_U\}$, où U est le nombre de catégories uniques.
2. **Création d'un vecteur binaire pour chaque catégorie** : pour chaque catégorie g_h, il y'a création d'un vecteur de longueur U. Dans ce vecteur, la h-ème position est 1 (indiquant la présence de la catégorie g_h) et toutes les autres positions sont égales à 0.
3. **Transformation des données** : chaque valeur catégorielle de x_j est remplacée par son vecteur binaire correspondant.

Soit une variable catégorielle x_j avec U catégories uniques $\{g_1, g_2, \ldots, g_U\}$. Chaque catégorie g_h est représentée par un vecteur binaire V_h de longueur U, tel que :

$$V_h = [0, 0, \ldots, 1, \ldots, 0]$$

Dans ce vecteur, la h-ème position est égale à 1, et les autres positions sont égales à 0. Ce vecteur est appelé vecteur *one-hot* car un seul élément est "chaud" (1) et les autres sont "froids" (0).

3.4.9 Sélection des variables

La sélection de variables explicatives basée sur la corrélation consiste à analyser les relations linéaires entre les variables explicatives et l'état de réservation, ainsi qu'entre les variables explicatives elles-mêmes. Le processus commence par le calcul du coefficient de corrélation, en utilisant la corrélation de Pearson pour des variables continues dans l'équation 3.8, afin de quantifier la force et la direction de ces relations. Elle repose sur le calcul de la covariance entre deux variables continues. Le coefficient de corrélation est en fait la standardisation de la covariance. Cette standardisation permet d'obtenir une valeur qui variera toujours entre -1 et +1, peu importe l'échelle de mesure des variables mises en relation, où des valeurs proches de +1 indiquent une forte corrélation positive, des valeurs proches de -1 indiquent une forte corrélation négative, et des valeurs proches de 0 suggèrent l'absence de relation linéaire. La corrélation $Corr_{j_1,j_2}$ entre deux variables x_{j_1} et x_{j_2} avec n est le nombre total de données,

est définie comme suit :

$$Corr_{j_1,j_2} = \frac{n\sum_{i=1}^{n}(x_{i,j_1}x_{i,j_2}) - \sum_{i=1}^{n} x_{i,j_1}\sum_{i=1}^{n} x_{i,j_2}}{\sqrt{\left[n\sum_{i=1}^{n} x_{i,j_1}^2 - \left(\sum_{i=1}^{n} x_{i,j_1}\right)^2\right]\left[n\sum_{i=1}^{n} x_{i,j_2}^2 - \left(\sum_{i=1}^{n} x_{i,j_2}\right)^2\right]}}. \tag{3.8}$$

La matrice de corrélation entre toutes les variables numériques est présenté dans la figure 3.12.

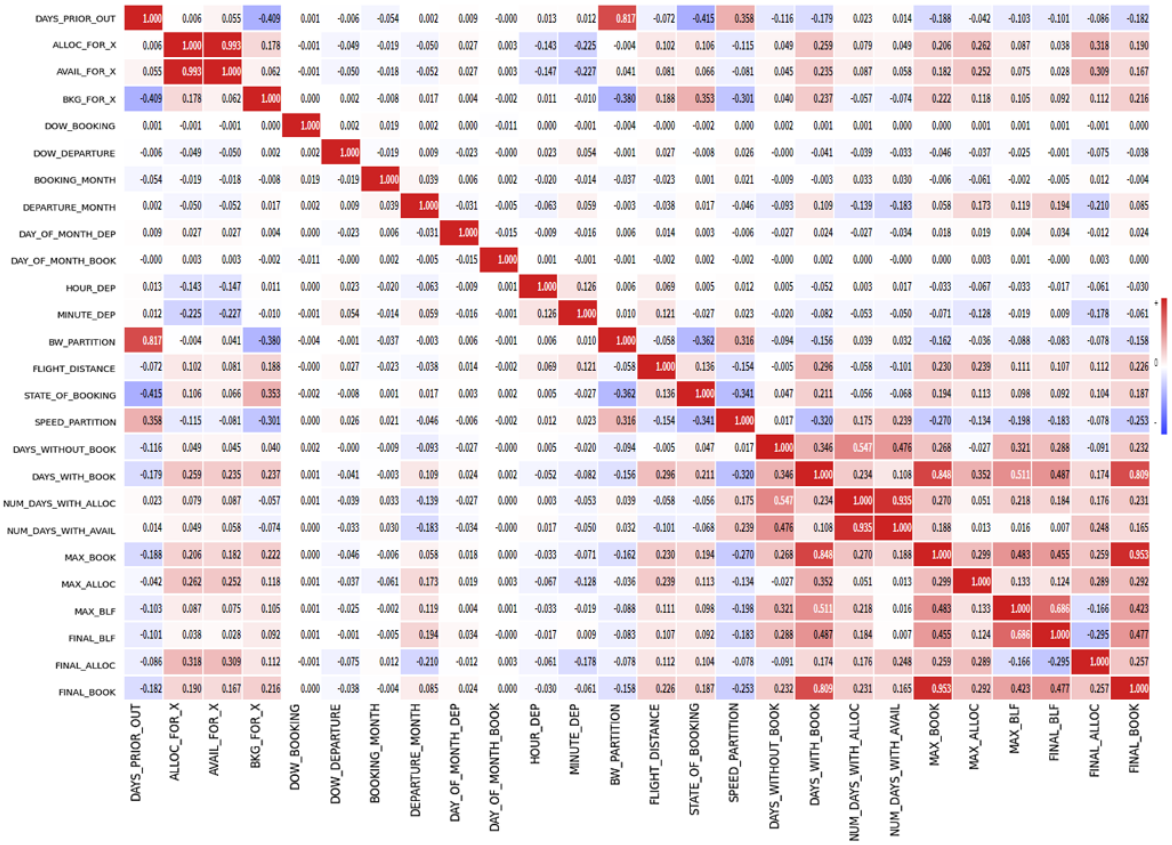


FIGURE 3.12 Matrice de corrélation

Dans le cadre de la sélection de variables, les corrélations entre les variables explicatives sont examinées pour détecter la multicolinéarité, c'est-à-dire lorsque deux ou plusieurs variables sont fortement corrélées entre elles. Dans ce cas, les variables redondantes sont supprimées afin de réduire la redondance et d'éviter le surapprentissage. On trouve cette situation entre Final_blf et Max_blf, Final_book et Max_book, Days_without_book et Days_with_book, Num_days_with_alloc et Num_days_with_avail, et Days_Prior_out et BW_partition, qui présentent une corrélation supérieure à 0,6. Des variables pour chaque paire sont éliminées.

Bien que les informations relatives à la date de départ, à la date de réservation ou au moment

du départ aient une faible corrélation avec la variable cible, elles ne sont pas exclues. Elles permettent de capter les effets saisonniers, les différences selon les jours de la semaine, ou les comportements spécifiques à certaines périodes.

En sélectionnant les variables en fonction de leur corrélation avec la variable cible et en éliminant celles qui sont redondantes, cette méthode simplifie le modèle et améliore son interprétabilité tout en conservant un pouvoir prédictif pertinent. Les variables explicatives retenues sont présentées dans le tableau 3.8. Il existe 22 variables.

Tableau 3.8 Variables explicatives retenues

Variable	Étiquette
Itinéraire	OD
Origine	Orig
Destination	Dest
Fenêtre de réservation	Days_Prior_out
Nombre d'allocations	Alloc_for_X
Jour de la semaine de départ	DOW_departure
Jour du mois de départ	day_of_month_dep
Mois de départ	Departure_month
Jour de la semaine de réservation	DOW_booking
Jour du mois de réservation	day_of_month_book
Mois de réservation	Booking_month
Heure du temps de départ	Hour_dep
Minute du temps de départ	Minute_dep
Distance du vol	Flight_distance
Région	Region_code
Nombre maximale d'allocations par vol	Max_alloc
Nombre final d'allocations par vol	Final_alloc
Nombre final de réservations par vol	Final_book
Taux de réservation final par vol	Final_blf
Nombre de jours sans réservation par vol	Days_without_book
Nombre de jours présentant d'allocations par vol	Num_days_with_alloc
Vitesse de réservation moyenne	Speed_partition

3.4.10 Division de l'ensemble des données par région

À ce stade, le jeu de données final destiné à la modélisation, noté D_f, a été constitué. Cet ensemble contient différentes variables x_j, présentées dans le tableau 3.8, où j varie de 1 à m, avec $m = 22$ représentant le nombre de variables explicatives. Cet ensemble est défini comme suit :

$$\{x_j \mid j = 1, 2, \ldots, m\}$$

En complément, la variable cible y, correspondant à l'état de réservation, est incluse. Comme l'état de réservation est modélisé par région, D_f est segmenté selon cette variable afin d'obtenir les ensembles présentés dans le tableau 3.9.

Tableau 3.9 Ensembles de données par région

Région	Ensemble de données	Nombre d'échantillons
Domestique	$D^{\text{domestique}}$	26 279 229
États-Unis	D^{usa}	16 865 371
Sud	D^{sud}	1 224 857
Pacifique	$D^{\text{pacifique}}$	618 829
Atlantique	$D^{\text{atlantique}}$	1 014 639

Ce chapitre a présenté en détail le processus de préparation des données nécessaires à la modélisation de la probabilité de réservation de surplus. À partir des ensembles D_{2023} et D_{2022}, une méthodologie a été appliquée pour structurer, nettoyer, enrichir et transformer les données brutes en un ensemble prêt à l'analyse prédictive. Les variables explicatives ont été sélectionnées en fonction de leur pertinence statistique et de leur capacité à représenter les dynamiques de réservation, tout en tenant compte des enjeux de multicolinéarité et de distribution. Des transformations spécifiques, telles que l'encodage des variables catégorielles et la réduction de l'asymétrie par la fonction `log1p`, ont permis d'optimiser les données pour leur intégration dans les modèles d'apprentissage automatique. Le jeu de données final, noté D_f, a été segmenté par région afin d'adapter les modèles aux spécificités de chaque marché. Ce cadre de préparation assure la fiabilité et la robustesse des analyses effectuées dans la phase de modélisation et d'interprétation des résultats.

CHAPITRE 4 MÉTHODOLOGIES DE MODÉLISATION

Dans ce chapitre, les différentes stratégies de modélisation de l'état de réservation sont présentées à l'aide des ensembles de données préparés pour les différentes régions. Trois stratégies de modélisation sont appliquées afin d'améliorer les résultats de la classification. La première stratégie illustre le processus de modélisation de base et permet de confirmer le choix du modèle d'apprentissage automatique, LightGBM, tel que présenté dans la section 4.1. La deuxième stratégie repose sur la création de groupes de données selon les partitions de fenêtres de réservation, sur lesquels la modélisation est effectuée, comme expliqué dans la section 4.2. La troisième stratégie ajoute une seconde variable pour la création des groupes, à savoir le mois de départ, comme détaillé dans la section 4.3. Ces deux dernières stratégies visent à améliorer les performances de modélisation.

Par ailleurs, la méthodologie utilisée pour analyser l'effet des différents facteurs sur la probabilité de réservation de sièges de surplus dans les sept jours à venir est également présentée dans la section 4.4.

4.1 Méthodologie de la première stratégie

Le processus de modélisation de la première stratégie est illustré dans la figure 4.1. Il a été appliqué sur les données relatives à chaque région. Le processus commence par la division de l'ensemble des données en un ensemble d'apprentissage et un ensemble de test, tel expliqué dans la section 4.1.1. Ensuite, le problème des données déséquilibrées est résolu en utilisant soit la technique de suréchantillonnage synthétique des minorités (SMOTE), soit la pondération des données, comme expliqué dans la section 4.1.2. Le processus se poursuit par l'entraînement du modèle LightGBM sur les données d'apprentissage, comme décrit à la section 4.1.3, ce qui aboutit à la création d'un modèle entraîné. Ce modèle est ensuite testé par rapport à des données de test, qui représentent de nouvelles informations pour le modèle.

Enfin, la phase de construction du modèle se termine par l'évaluation de sa performance à l'aide de diverses métriques, détaillé à la section 4.1.4. Une étude comparative entre le modèle LightGBM et d'autres modèles d'apprentissage automatique est expliquée à la section 4.1.5, suivie d'une comparaison avec un modèle de référence, expliquée à la section 4.1.6, afin de confirmer le choix de ce modèle et de s'assurer qu'il offre la meilleure performance.

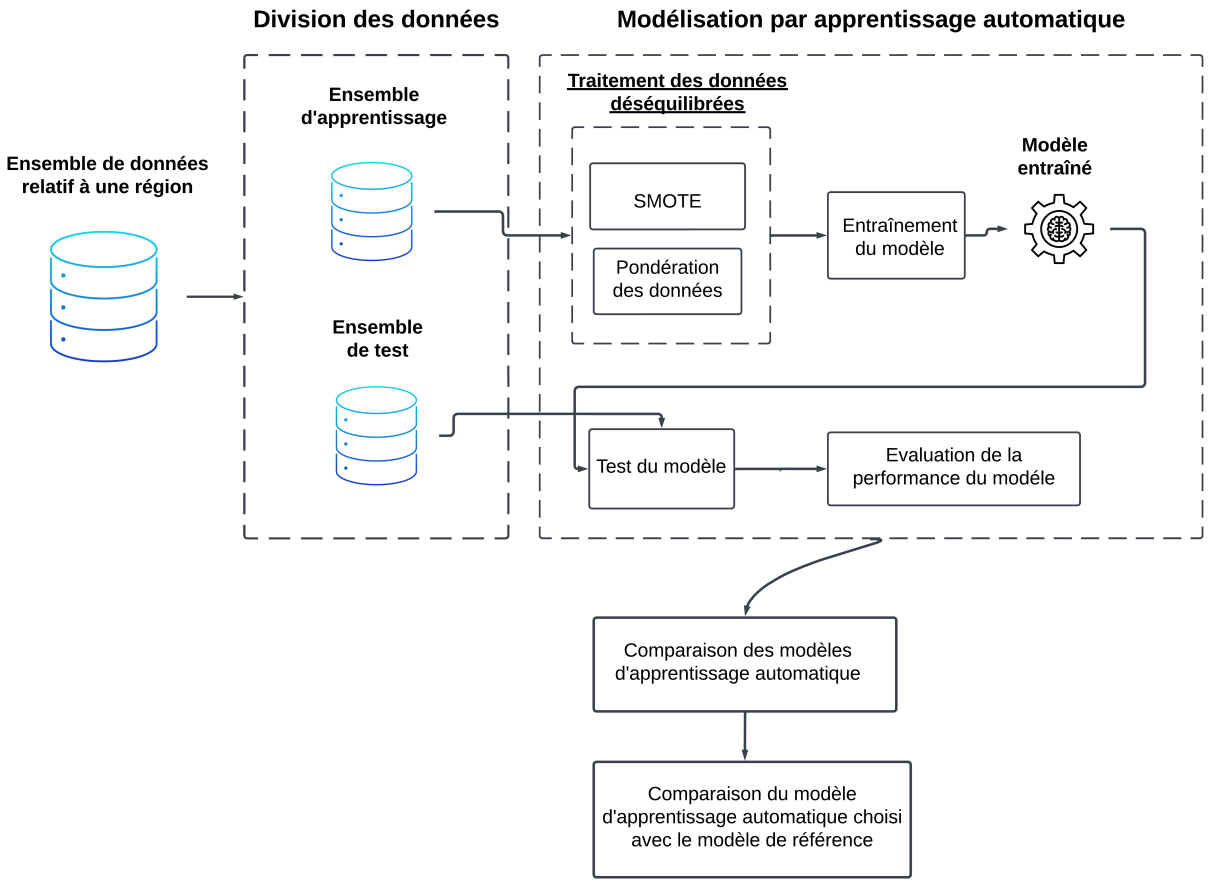


FIGURE 4.1 Méthodologie de la première stratégie

4.1.1 Division des données

Pour chaque ensemble de données relatif à une région, la répartition entre l'ensemble d'apprentissage et l'ensemble de test a été réalisée selon un ratio de 80%/20% de manière aléatoire. Cela signifie que 80% des données sont utilisées pour entraîner le modèle, tandis que 20% sont réservées pour tester ses performances sur des données non vues. Cependant, afin de préserver la distribution de la variable cible, l'état de réservation, un échantillonnage stratifié a été appliqué. Cet échantillonnage garantit que les proportions des différentes classes de la variable cible sont maintenues à la fois dans les ensembles d'apprentissage et de test.

Par ailleurs, une validation croisée à 5 plis est utilisée pour garantir la robustesse du modèle et améliorer sa capacité de généralisation. La figure 4.2 illustre ce processus : l'ensemble des données d'apprentissage est divisé en cinq plis de taille équivalente. À chaque itération, quatre plis sont utilisés pour l'entraînement et le pli restant pour la validation. Ce processus est répété cinq fois, en alternant les rôles des plis, de manière à ce que chaque pli soit utilisé

une fois comme jeu de validation.

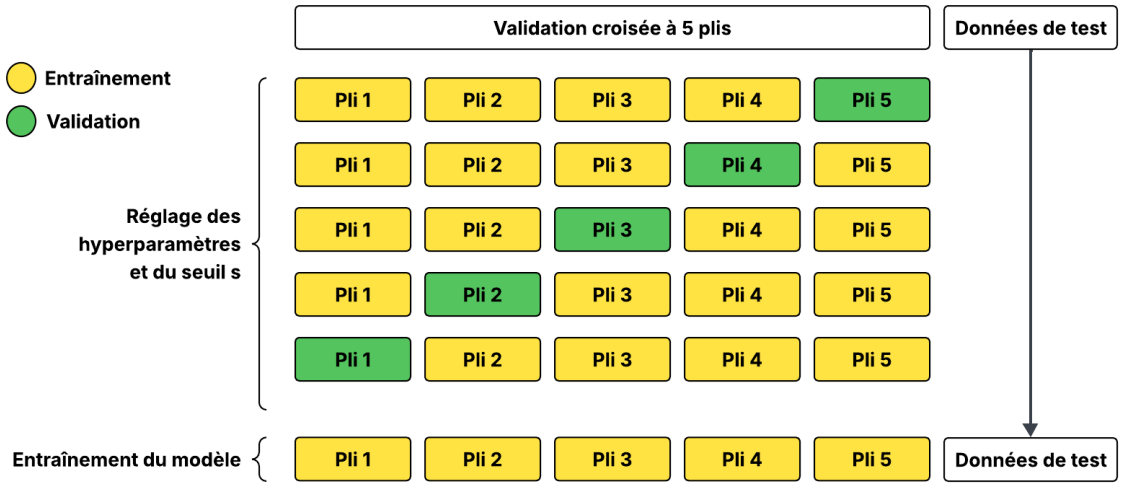


FIGURE 4.2 Validation croisée à 5 plis

La validation croisée sert au réglage des hyperparamètres du modèle comme le taux d'apprentissage, le nombre de feuilles et la profondeur maximale des arbres. Pour chaque combinaison d'hyperparamètres, la performance du modèle est évalué sur les cinq validations, et la configuration offrant les meilleurs résultats moyens est retenue. Cette procédure permet de limiter le surapprentissage.

En outre, le seuil de classification s est déterminé à partir des résultats obtenus pendant la validation croisée. Bien que le seuil par défaut soit fixé à 0,5, il est ajusté dans le but d'optimiser le F1-score. Le seuil optimal est sélectionné sur les prédictions des plis de validation uniquement, puis appliqué au modèle final pour convertir les probabilités en classes sur l'ensemble de test.

Enfin, un modèle final est entraîné sur l'ensemble complet des données d'apprentissage, c'est-à-dire les cinq plis combinés. Ce modèle final est ensuite utilisé pour effectuer des prédictions sur un ensemble de test indépendant.

4.1.2 Résolution du problème des données déséquilibrées

Deux méthodes ont été essayées pour résoudre le problème de données déséquilibrées, la première utilisant la pondération des données et SMOTE étant la deuxième technique.

Pondération des données

La pondération pour les données déséquilibrées, d'après Feizi et al. [75], est une technique utilisée en modélisation pour résoudre le problème de la représentation disproportionnée des différentes classes de l'état de réservation. Les ensembles de données pour les différentes régions montrent un déséquilibre où la classe majoritaire indiquant qu'il n'y a pas de réservation de surplus dans les sept prochains jours est plus présente que l'autre classe.

Cela conduit à un entraînement de modèle biaisé et à des prédictions inexactes, car les modèles ont tendance à favoriser la classe majoritaire. Pour contrer ce biais, la pondération attribue une importance accrue aux échantillons de classe minoritaire lors de l'entraînement du modèle, leur donnant plus d'influence dans le processus d'apprentissage. Cela est généralement réalisé en attribuant des poids plus élevés aux échantillons de la classe minoritaire par rapport à celles de la classe majoritaire. Les poids sont déterminés en fonction de la distribution des classes, dans le but de rendre le modèle plus sensible à la classe minoritaire sans totalement négliger la classe majoritaire. Ce faisant, la pondération contribue à obtenir des performances plus équilibrées sur toutes les classes, améliorant la capacité du modèle à généraliser correctement sur des données non observées et à améliorer sa précision prédictive.

Dans LightGBM, lorsqu'il s'agit de déséquilibre des classes, la fonction de perte standard est modifiée pour incorporer des poids qui sont inversement proportionnels aux fréquences des classes selon Han et al. [76]. La fonction de perte pondérée de l'entropie croisée binaire présentée dans l'équation 2.2 devient comme suit :

$$L = -\sum_{i=1}^{n_{train}} \lambda_i \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right], \tag{4.1}$$

où λ_i est le poids attribué. Il est déterminé en fonction de la distribution des classes, dans le but de donner un poids plus élevé aux échantillons de classe minoritaire.

Pendant le processus d'optimisation, LightGBM minimise cette fonction de perte pondérée par des gradients, où ils sont calculés par rapport aux paramètres du modèle et utilisés pour le mettre à jour à chaque itération. Les poids attribués aux échantillons influencent l'amplitude des gradients, garantissant que le modèle apprend efficacement à partir des échantillons des classes majoritaires et minoritaires.

Technique de suréchantillonnage synthétique des minorités (SMOTE)

SMOTE est une autre méthode essayée pour résoudre le problème des ensembles de données déséquilibrés en apprentissage automatique selon Nguyen et al. [77]. Il fonctionne comme

suit :

1. **Identification de la classe minoritaire** : SMOTE se concentre sur la classe indiquant l'existence de réservation de surplus, qui est sous-représentée par l'autre classe.
2. **Génération d'échantillons synthétiques** :
   — Au lieu de simplement dupliquer les échantillons de la classe minoritaire, Il crée des échantillons synthétiques en sélectionnant deux ou plusieurs échantillons similaires de cette classe et en créant de nouvelles échantillons interpolées entre elles.
   — Pour chaque échantillon de la classe minoritaire, SMOTE identifie ses k plus proches voisins en utilisant la distance euclidienne. Il sélectionne ensuite aléatoirement l'un de ces voisins et crée un nouvel échantillon synthétique entre l'échantillon d'origine et son voisin.
3. **Équilibrage de l'ensemble de données** : En ajoutant ces nouveaux échantillons synthétiques, SMOTE augmente la représentation de la classe indiquant l'existence de réservation de surplus, rendant l'ensemble de données plus équilibré et permettant au modèle d'apprendre à mieux classifier l'état de réservation.

SMOTE réduit le biais des modèles en faveur de la classe majoritaire en fournissant des données d'apprentissage plus équilibrées. Il prévient aussi le surapprentissage qui peut survenir lors de la duplication des échantillons de la classe minoritaire.

4.1.3 Entraînement du modèle LightGBM

LightGBM construit un ensemble d'arbres de décision de manière séquentielle, chaque arbre étant conçu pour corriger les erreurs des arbres précédents. Ce processus s'appuie sur l'algorithme décrit dans la section 2.4.2 pour estimer la probabilité de réservation de surplus dans les sept jours à venir et générer les prédictions concernant l'état de réservation. Les caractéristiques spécifiques de ce modèle sont présentées dans la section 2.4.1.

Dans cette section, nous nous concentrons sur la création de ces arbres en utilisant un exemple afin de mieux comprendre la construction du modèle. La structure de l'arbre de décision dans LightGBM est composée de nœuds représentant des divisions des variables explicatives et de nœuds feuilles contenant les prédictions. À chaque division, le modèle optimise une fonction objective spécifique selon Truong et al. [78]. Il prend en charge divers critères de division ainsi que des techniques d'élagage pour améliorer les performances du modèle. Une fois le premier arbre créé, les autres arbres sont construits de manière récursive selon une approche feuille par feuille, en maximisant la réduction de la perte à chaque étape de l'apprentissage selon Li et al. [79]. Ces arbres forment collectivement un modèle d'ensemble, où les prédictions

sont obtenues en agrégeant les résultats des arbres individuels, aboutissant ainsi à un modèle d'apprentissage automatique efficace.

Pour mieux comprendre le fonctionnement de ce modèle, un exemple est présenté en utilisant une partie des données relatives aux vols au départ en janvier dans la région Domestique. Cela représente un total de $n = 4700$ échantillons, répartis en $n_{\text{train}} = 3760$ échantillons dédiés à l'apprentissage du modèle et $n_{\text{test}} = 940$ échantillons réservés au test.

La structure du modèle LightGBM dans cet exemple contient $M = 3$ arbres de décision, et le nombre maximal de feuilles est 6 dans un arbre unique, ce qui constitue le critère d'arrêt. De plus, la méthode SMOTE a été appliquée afin d'obtenir une distribution parfaitement équilibrée entre les classes. Cela implique que la valeur initiale de log-odds est nulle pour tous les échantillons, c'est-à-dire $z_i^{(0)} = 0$. En outre, aucun rétrécissement d'apprentissage n'est appliqué, $\alpha = 1$, en raison du faible nombre d'arbres et de la taille réduite du jeu de données, ce qui limite le risque de surapprentissage. Cette configuration permet également de simplifier l'interprétation du processus d'apprentissage en rendant les contributions de chaque arbre directement observables.

L'apprentissage du modèle commence par la création du premier arbre de décision, illustré dans la figure 4.3. Le processus suit les étapes suivantes :

1. **Calcul des gradients** : Le modèle calcule le gradient négatif de la fonction de perte, présenté dans l'équation 2.3, pour tous les échantillons. Ce gradient, qui est le pseudo-résidu détaillé dans l'équation 2.9, représente l'impact de chaque échantillon sur la perte totale et indique la direction dans laquelle le modèle ajuste ses paramètres afin de minimiser cette perte.
2. **Choix de la meilleure division** : À chaque nœud, le modèle recherche la meilleure combinaison entre la variable explicative x_j et le seuil de séparation Θ afin de définir deux régions :
$$\{x \mid x_j < \Theta\}, \quad \{x \mid x_j > \Theta\} \tag{4.2}$$
La meilleure division est déterminée sur la base du gain de division. Le modèle sélectionne la combinaison de x_j et Θ qui maximise ce gain, garantissant ainsi une séparation optimale des données.
3. **Calcul du gain de division** : Le gain de division est calculé comme suit :
$$\text{Gain} = \frac{(L_{\text{sum}})^2}{L_{\text{count}} + \rho} + \frac{(R_{\text{sum}})^2}{R_{\text{count}} + \rho} - \frac{(T_{\text{sum}})^2}{T_{\text{count}} + \rho} - \epsilon. \tag{4.3}$$
où :

— $L_{\text{sum}}, R_{\text{sum}}, T_{\text{sum}}$ sont les sommes des valeurs de gradient pour les échantillons à gauche, à droite et au total.

— $L_{\text{count}}, R_{\text{count}}, T_{\text{count}}$ sont le nombre d'échantillons dans chaque division.

— ρ est le terme de régularisation pour éviter le sur-apprentissage, fixé à 0.2.

— ϵ est le gain minimum pour diviser (terme d'élagage), fixé à 0.1.

Le choix du prochain nœud à diviser repose également sur le gain de division. Le nœud présentant le gain le plus élevé est divisé en priorité, ce qui définit la croissance de l'arbre selon une approche feuille par feuille.

4. **Répétition des étapes 2 et 3** : Le modèle continue à diviser en répétant les étapes 2 et 3 jusqu'à ce que le critère d'arrêt soit atteint, à savoir le nombre maximal de feuilles.

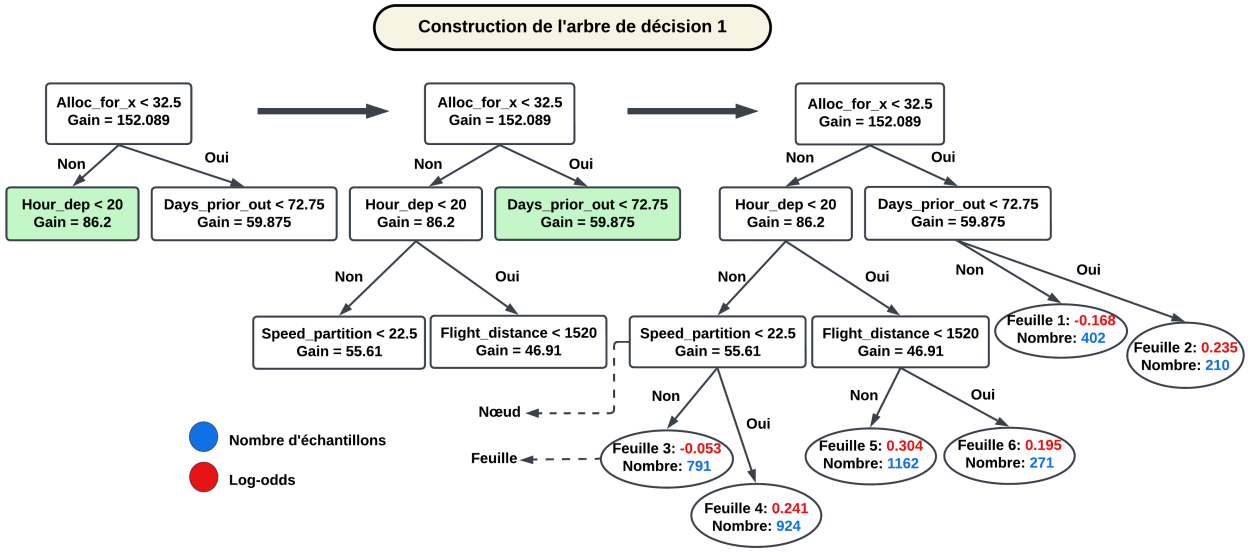


FIGURE 4.3 Construction du premier arbre de décision de LightGBM

Dans l'exemple utilisé, la première division est effectuée sur le nombre d'allocations avec $\Theta = 32.5$, entraînant un gain de 152.089. Cette division génère deux nouveaux nœuds : fenêtre de réservation < 72.75 et heure de départ < 20. Dans ce cas, le premier nœud présente un gain plus élevé que le second ($86.2 > 59.875$), donc heure de départ < 20 est le prochain nœud divisé, créant ainsi deux nouveaux nœuds : vitesse de réservation moyenne < 22.5 et distance de vol < 1520. À ce stade, le nœud fenêtre de réservation < 72.75 présente le gain le plus élevé, soit 59.875. Comme le nombre maximal de feuilles est fixé à 6, les deux premières feuilles sont créées. D'autres feuilles sont ensuite formées à partir des deux nœuds restants jusqu'à atteindre 6 feuilles. Chaque feuille contient le nombre d'échantillons repré-

sentés dans l'ensemble de données ainsi que la contribution en log-odds, utilisées pour estimer la probabilité de réserver de surplus dans les sept jours suivants.

Les autres arbres sont ensuite construits de manière séquentielle, où chaque nouvel arbre est ajusté pour minimiser la perte résiduelle, en apprenant aux gradients négatifs selon Sheridan et al. [80]. Cela est illustré dans la figure 4.4, indiquant l'écart entre les prédictions du modèle et les valeurs réelles selon Zhang et al. [81]. Ces gradients permettent d'orienter l'arbre vers les zones les plus critiques à améliorer. Après l'apprentissage de chaque arbre, ses prédictions sont ajustées par un taux d'apprentissage et ajoutées au modèle précédent, affinant progressivement la prédiction globale selon Li et al. [82]. Ce processus se poursuit jusqu'à atteindre le nombre d'arbres fixé, soit $M = 3$.

Par ailleurs, la validation croisée a été appliquée à l'ensemble d'apprentissage dont le seuil de classification s est optimisé en maximisant une mesure de performance spécifique, le F1-score, qui concilie la précision et le rappel. Pour ce faire, différentes valeurs du seuil s sont testées ; pour chacune, le F1-score est calculé, puis la valeur de s maximisant ce score moyen est sélectionnée. Le seuil de 0,6 a ainsi été identifié comme optimal et retenu pour la prise de décision finale.

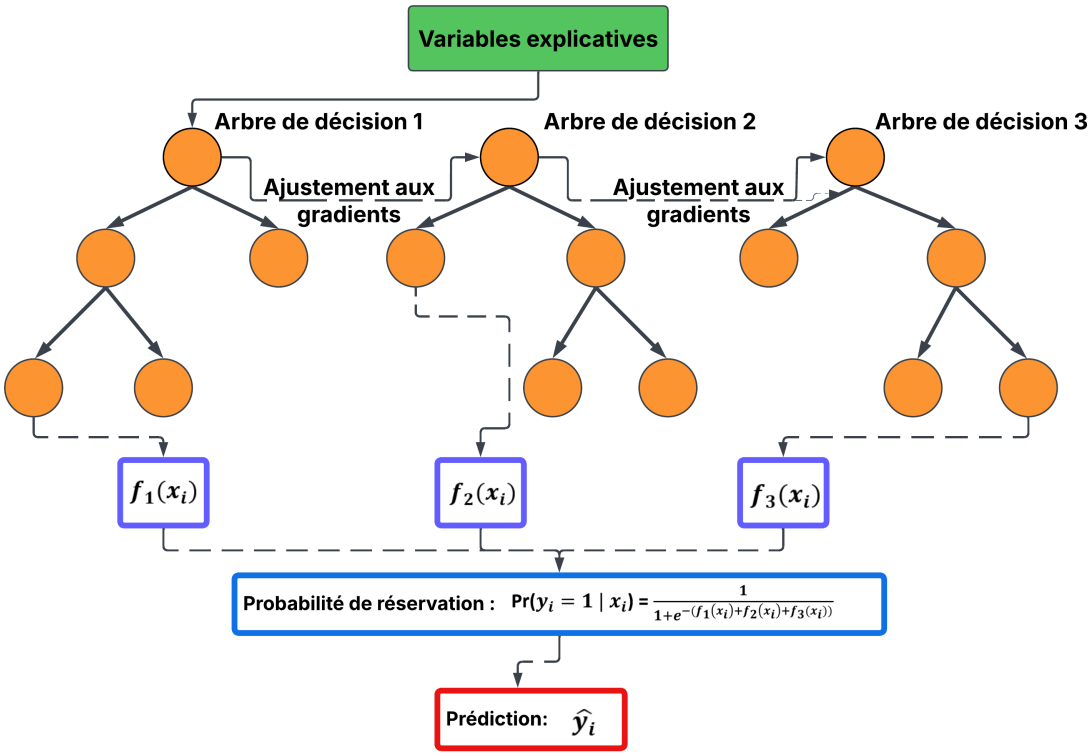


FIGURE 4.4 Construction d'arbres de décision LighGBM

Après l'entraînement du modèle LightGBM, l'estimation de la probabilité de réserver de surplus dans les sept prochains jours est déterminée, en se basant sur quelques échantillons de l'ensemble de test présentés dans le tableau 4.1, et leurs prédictions sont conclues.

Tableau 4.1 Échantillons de test

Échantillon	Alloc_for_x	Hour_dep	Days_prior_out	Speed_partition	Flight distance	Autres variables	Reservation_status
x_1	48	22	62	56	445	...	0
x_2	21	8	49	37	619	...	1
x_3	35	10	35	19	420	...	1

Pour chaque arbre, la contribution en log-odds de la feuille à laquelle appartient l'échantillon est collectée. Le processus est le suivant :

1. Commencer à la racine de l'arbre.
2. Suivre les règles de décision à chaque nœud en fonction des variables explicatives de l'échantillon.
3. Arriver à un nœud feuille et récupérer la valeur de la feuille, qui correspond à la contribution en log-odds $f_w(x_i)$ de l'arbre w à la prédiction finale en log-odds pour i-ème échantillon.

Par exemple, dans le premier arbre, le premier échantillon appartient à la feuille 3 avec une contribution en log-odds de -0.053, le deuxième échantillon appartient à la feuille 2 avec une contribution en log-odds de 0.235 et le troisième échantillon appartient à la feuille 6 avec une contribution en log-odds de 0.195. Le modèle collecte les mêmes informations pour les autres arbres. La probabilité de réserver de surplus dans les sept prochains jours est montrée dans l'équation 2.15.

Pour déterminer la classe de l'état de réservation, qui constitue le second résultat du modèle, après avoir obtenu l'estimation de probabilité, le modèle utilise le seuil de classification s. Il définit le seuil de probabilité à partir duquel un échantillon est classé en classe minoritaire, indiquant l'existence de la réservation de surplus. Mathématiquement, cette règle de décision est exprimée dans l'équation 2.16.

Les probabilités résultantes et les prédictions des trois échantillons sont présentées dans le tableau 4.2.

Après avoir expliqué le fonctionnement du modèle utilisé dans cette recherche, On l'applique pour modéliser la totalité des données, où chaque modèle présente $M = 10$ arbres de décision.

Tableau 4.2 Résultats des probabilités de réservation de surplus et des prédictions

Résultat	x_1	x_2	x_3
$f_1(x_i)$	-0.053	0.235	0.195
$f_2(x_i)$	-0.127	0.513	0.164
$f_3(x_i)$	0.008	0.171	0.437
p_i	45.7%	71.5%	68.9%
$\hat{y}_i$	0	1	1

4.1.4 Évaluation de la performance du modèle

La performance de la modélisation est évaluée à l'aide du F1-score, du rappel, de la précision et de l'exactitude. Ces mesures sont calculées à partir des prédictions obtenues sur l'ensemble de test, ce qui permet d'évaluer la capacité de généralisation du modèle. Avant la présentation de ces mesures, quelques notions essentielles sont définies.

- **TP (Vrai Positif)** : Nombre de cas où le modèle prédit qu'il y a de réservation de surplus dans les sept prochains jours (1) et qu'il y a effectivement une réservation de surplus.
- **TN (Vrai Négatif)** : Nombre de cas où le modèle prédit qu'il n'y a pas de réservation de surplus dans les sept prochains jours (0) et qu'il n'y a effectivement pas de réservation de surplus.
- **FP (Faux Positif)** : Nombre de cas où le modèle prédit qu'il y a de réservation de surplus dans les sept prochains jours (1), alors qu'il n'y en a pas effectivement.
- **FN (Faux Négatif)** : Nombre de cas où le modèle prédit qu'il n'y a pas de réservation de surplus dans les sept prochains jours (0), alors qu'il y en a effectivement une.

Ensuite, différentes mesures utilisées pour évaluer la performance d'un modèle d'apprentissage automatique.

Précision : rapport entre le nombre de vrais positifs et le nombre total de prédictions positives qui est égal à la somme des vrais positifs et des faux positifs. Elle mesure la proportion de prédictions positives correctes parmi toutes les prédictions positives effectuées par le modèle.

$$\text{Précision} = \frac{TP}{TP + FP}. \tag{4.4}$$

Rappel : appelé sensibilité ou taux de vrais positifs, est le rapport entre le nombre de vrais positifs et le nombre total de réels positifs qui est égal à la somme des vrais positifs et des faux négatifs. Il mesure la capacité du modèle à identifier correctement toutes les échantillons positives.

$$\text{Rappel} = \frac{TP}{TP + FN}. \tag{4.5}$$

Exactitude : rapport entre le nombre de prédictions correctes qui est égal à la somme des vrais positifs et des vrais négatifs et le nombre total d'échantillons évaluées. Elle mesure la proportion d'échantillons correctement classées parmi toutes les échantillons.

$$\text{Exactitude} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.6}$$

F1-score : utilisé pour évaluer l'équilibre entre la précision et le rappel. Il combine ces deux métriques en un seul indicateur, particulièrement utile lorsque les classes sont déséquilibrées. Il cherche à trouver un équilibre entre la précision et le rappel.

$$F_1 = 2 \cdot \frac{\text{Précision} \cdot \text{Rappel}}{\text{Précision} + \text{Rappel}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}. \tag{4.7}$$

4.1.5 Comparaison des modèles d'apprentissage automatique

Dans cette étape, nous définissons les modèles d'apprentissage automatique utilisés pour comparer ses performances avec celles de LightGBM dans la prédiction de l'état de réservation.

- **k plus proches voisins (k-NN)** : est un algorithme de classification basé sur la proximité entre les points de données. Lorsqu'un nouvel échantillon doit être classé, l'algorithme identifie les k voisins les plus proches de cet échantillon dans l'espace des variables explicatives selon Wu et al. [83]. La classification est ensuite déterminée en fonction de la classe majoritaire parmi ces voisins. Le modèle utilise des mesures de distance, comme la distance Euclidienne, pour déterminer les voisins. Il n'a pas de phase d'entraînement spécifique, mais effectue des calculs de proximité chaque fois qu'il doit prédire une classe. Le choix de k, le nombre de voisins à prendre en compte, a un impact sur la performance, car un k trop petit peut rendre le modèle sensible aux bruits, tandis qu'un k trop grand peut le rendre trop général.
- **Régression Logistique** : est un modèle de classification qui estime la probabilité qu'un échantillon appartienne à une classe donnée, en utilisant une fonction logistique sigmoïde. Ce modèle, d'après Ma [84] applique une combinaison linéaire des variables explicatives, puis passe cette combinaison dans une fonction sigmoïde pour obtenir une probabilité comprise entre 0 et 1. En fonction de cette probabilité, l'échantillon est classé dans une classe. La régression logistique ajuste les coefficients des caractéristiques durant la phase d'entraînement pour minimiser l'erreur entre les prédictions

et les valeurs réelles, en utilisant des techniques comme la descente de gradient.

— **Forêt Aléatoire** : est un modèle d'ensemble qui combine plusieurs arbres de décision pour améliorer la robustesse des prédictions. Chaque arbre est construit à partir d'un sous-ensemble aléatoire des données d'entraînement et des variables explicatives, ce qui réduit le risque de surapprentissage par rapport à un seul arbre selon Kirasich et al. [85]. Lorsque le modèle doit faire une prédiction, il génère une prédiction à partir de chaque arbre de la forêt, puis la classe finale est déterminée par la majorité des votes des arbres.

— **Arbre de Décision** : est un modèle de classification qui divisent l'espace des données en fonction des variables explicatives les plus pertinentes pour séparer les classes suivant Huang [86]. Lors de l'entraînement, l'algorithme crée un arbre en choisissant, à chaque nœud, la caractéristique qui permet de mieux séparer les données en fonction de critères comme l'entropie ou le gain d'information. Chaque branche de l'arbre représente une règle de décision basée sur une variable explicative spécifique, et l'échantillon suit ces règles jusqu'à arriver à une feuille, qui détermine la classe de l'échantillon.

4.1.6 Comparaison du modèle d'apprentissage automatique choisi avec un modèle de référence

Dans cette section, un modèle de référence est construit à l'aide de l'ensemble de données de test. L'objectif est d'évaluer la performance du modèle d'apprentissage automatique sélectionné en comparant ses métriques avec celles obtenues par le modèle de référence, afin de confirmer la pertinence du recours à l'apprentissage automatique.

Le modèle de référence est élaboré à partir d'une ingénierie des caractéristiques basée sur la variable du nombre de réservations incrémentale. L'état de réservation pour les sept jours suivants est prédit en fonction de la présence d'au moins une réservation de surplus au cours des sept jours précédents.

Il est supposé que, pour un vol donné, x_r désigne la variable du nombre de réservations incrémentale, et que e représente un indice d'échelonnement de la fenêtre de réservation, allant en ordre décroissant de 364 à 0 jours, où 0 correspond à la date de départ et 364 à la première date de publication. La valeur prédite de l'état de réservation $\hat{y_e}$, selon le modèle de référence, à e jours avant la date de départ, est définie comme suit :

$$\hat{y_e} = \begin{cases} 1 & \text{si } \sum_{o=1}^{7} x_{r,e+o} > 0, \\ 0 & \text{sinon.} \end{cases} \tag{4.8}$$

4.2 Méthodologie de la deuxième stratégie

La deuxième stratégie vise à améliorer les performances de modélisation. Pour chaque région, au lieu d'utiliser l'ensemble total des données pour entraîner le modèle LightGBM, des groupes de données sont crées en se basant sur une partition de la fenêtre de réservation, comme montré dans la figure 4.5. Ce regroupement permet de diviser les données en trois sous-ensembles définis comme suit :

— **Sous-ensemble 1** : Contient les données dont la fenêtre de réservation est comprise entre 0 et 120 jours, noté $0 \leq BW \leq 120$.
— **Sous-ensemble 2** : Contient les données dont la fenêtre de réservation est comprise entre 121 et 240 jours, noté $121 \leq BW \leq 240$.
— **Sous-ensemble 3** : Contient les données dont la fenêtre de réservation est comprise entre 241 et 364 jours, noté $241 \leq BW \leq 364$.

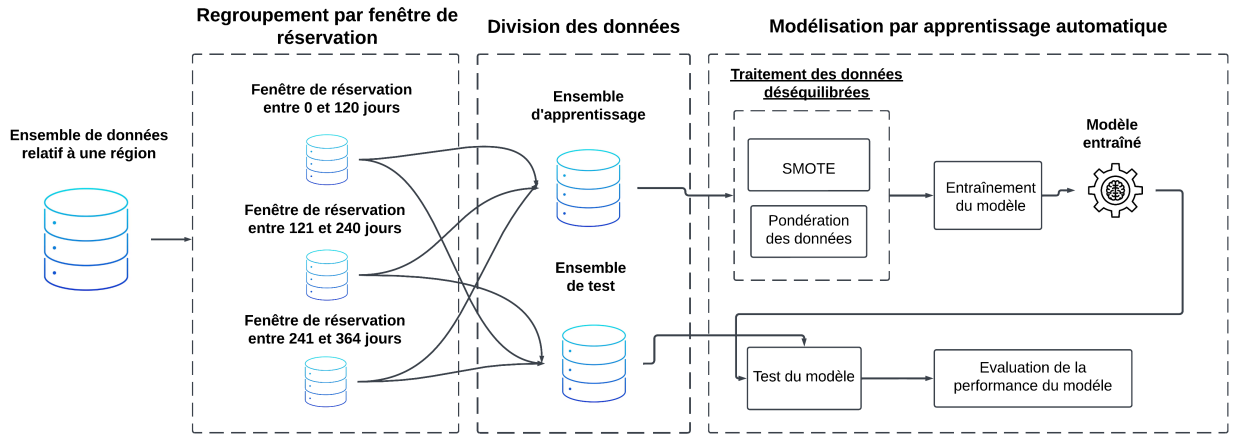


FIGURE 4.5 Méthodologie de la deuxième stratégie

Ensuite, chaque sous-ensemble est divisé en un ensemble d'apprentissage et un ensemble de test afin de modéliser l'état de réservation, en suivant le même processus décrit dans la section 4.1.3.

À la fin de cette étape, trois modèles entraînés sont obtenus. Leur performance est évaluée à l'aide des mêmes métriques décrites dans la section 4.1.4. Les mesures de performance des trois modèles sont ensuite agrégés en calculant leur moyenne, afin de les comparer aux résultats de la première stratégie.

4.3 Méthodologie de la troisième stratégie

La troisième stratégie a pour objectif d'améliorer davantage les performances de modélisation. Elle repose sur la deuxième stratégie, qui consiste à créer des sous-ensembles en fonction de la fenêtre de réservation pour chaque région. Après ce premier regroupement, un deuxième regroupement est effectué, cette fois en fonction du mois de départ. Chaque sous-ensemble ainsi formé correspond à un mois spécifique comme montré dans la figure 4.6. Cela donne un total de 3×12 sous-ensembles par région.

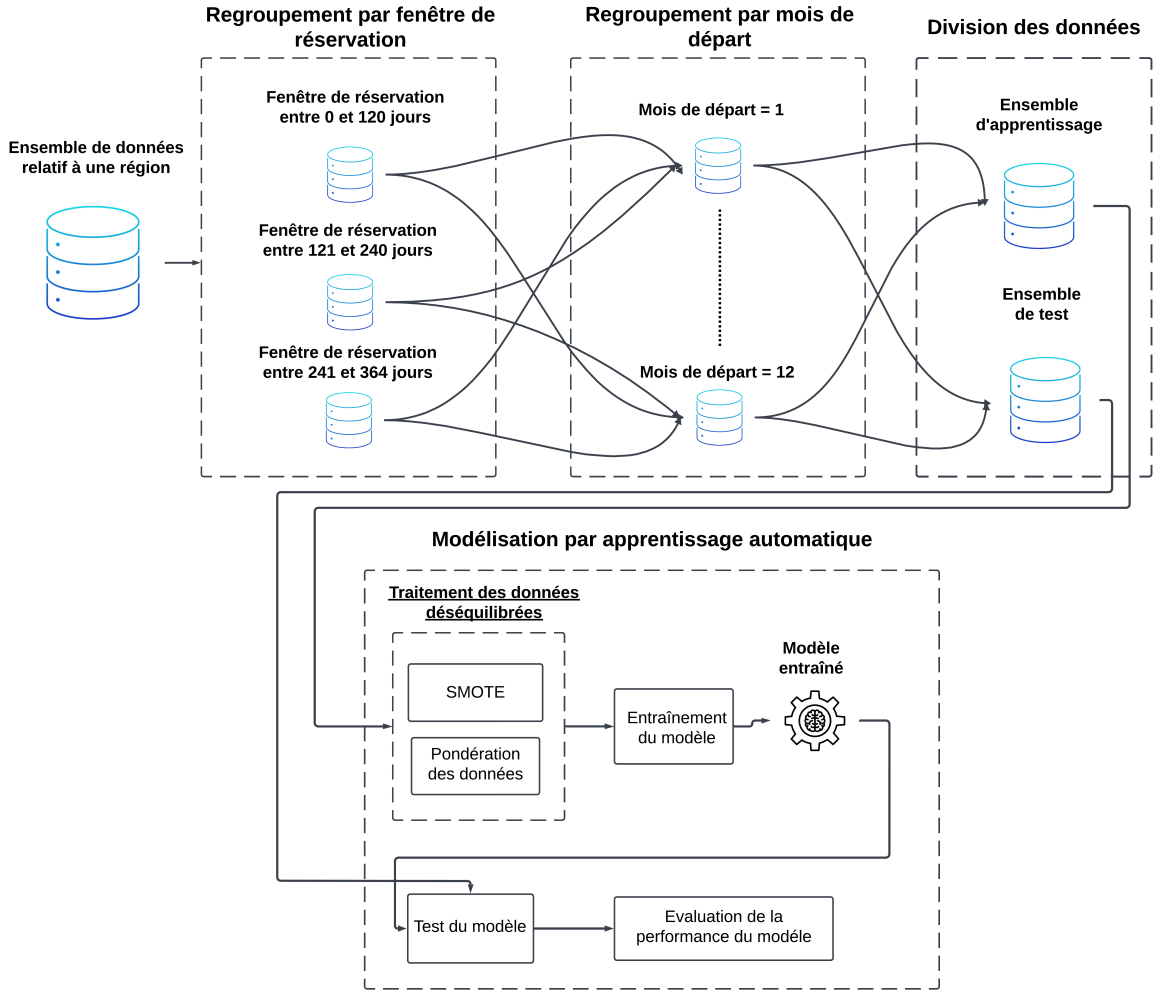


FIGURE 4.6 Méthodologie de la troisième stratégie

Ensuite, chaque sous-ensemble est divisé en un ensemble d'apprentissage et un ensemble de test, afin de modéliser l'état de réservation en suivant le même processus décrit dans la section 4.1.3.

À l'issue de cette étape, nous obtenons 3×12 modèles entraînés pour chaque région. Leur

performance est évaluée à l'aide des mêmes métriques détaillées dans la section 4.1.4. Les mesures de performance des 36 modèles sont ensuite agrégés en calculant leur moyenne, ce qui permet de les comparer aux résultats obtenus dans la deuxième stratégie.

4.4 Méthodologie utilisée pour comprendre l'effet des facteurs sur la probabilité de réservation de surplus

L'effet de différents facteurs sur la probabilité de réserver de surplus dans les sept prochains jours est évalué, en utilisant l'importance des variables à l'aide de l'impureté de Gini, expliquée dans la section 2.5.1. Ensuite, la dépendance partielle de certaines variables explicatives est déterminée pour observer l'effet indépendant de chacune d'entre elles sur la probabilité estimée. Cette technique est expliquée à la section 2.5.2.

CHAPITRE 5 RÉSULTATS DU PROJET

Dans ce chapitre, les résultats des différentes stratégies de modélisation visant à prédire l'état de réservation, sont présentés. Ces résultats permettent ensuite d'extraire les probabilités de réservation de surplus dans les sept jours à venir. L'importance des différentes variables explicatives dans la prédiction de l'état de réservation est également mise en évidence, ainsi que leur effet sur la probabilité estimée.

5.1 Résultats de la première stratégie de modélisation

Cette section a pour objectif de confirmer le choix du modèle d'apprentissage automatique, LightGBM. Ce modèle est comparé à d'autres modèles d'apprentissage automatique ainsi qu'au modèle de référence. Les mesures de performance du modèle LightGBM sont présentées dans la section suivante.

5.1.1 Performance du modèle LightGBM à prédire l'état de réservation

Dans cette section, les tableaux 5.1 et 5.2 montrent les mesures de performance du modèle LightGBM avant et après l'ajout des variables de qualité de vol, respectivement. Dans les deux cas, la méthode de pondération des données a été appliquée pour résoudre le déséquilibre des classes. Une amélioration est constatée sur l'ensemble des métriques, en particulier dans la région Domestique et aux États-Unis, où les données sont les plus déséquilibrées. Le F1-score a augmenté de 17% pour la région Domestique et de 15% pour les États-Unis. Pour les autres régions, bien que le gain de performance soit plus faible, les mesures sont supérieures à 70%, ce qui indique l'efficacité du modèle.

Tableau 5.1 Mesures de performance pour le modèle LightGBM par rapport à différentes régions sans ajout de variables de qualité de vol

Région	Exactitude	Précision	Rappel	F1-score
Domestique	86%	51%	44%	47%
États-Unis	86%	54%	52%	53%
Atlantique	87%	72%	69%	70%
Sud	90%	71%	69%	70%
Pacifique	91%	74%	68%	71%

Pour résoudre le problème des données déséquilibrées, une méthode de pondération des données a été utilisée, et ses résultats sont présentés dans le tableau 5.2. Dans le tableau 5.3,

Tableau 5.2 Mesures de performance pour le modèle LightGBM par rapport à différentes régions avec l'ajout de variables de qualité de vol

Région	Exactitude	Précision	Rappel	F1-score
Domestique	92%	64%	65%	64%
États-Unis	87%	69%	67%	68%
Atlantique	88%	74%	78%	76%
Sud	91%	75%	79%	77%
Pacifique	93%	76%	81%	79%

les résultats obtenus avec le modèle LightGBM en appliquant une autre technique, à savoir SMOTE, sont exposés. Ces deux techniques ont des impacts distincts sur les performances du modèle LightGBM. SMOTE permet une amélioration de la précision dans certaines régions, telles que l'Atlantique et le Sud, mais présente des performances inférieures en termes de rappel et de F1-score par rapport à la pondération. La pondération, en revanche, permet d'obtenir un meilleur rappel ainsi que des F1-score plus élevés, avec des performances plus équilibrées entre les régions. L'exactitude est également améliorée grâce à la pondération, en particulier dans les régions Domestique et Pacifique. En conclusion, SMOTE s'avère plus adapté à l'optimisation de la précision, tandis que la pondération permet une amélioration des mesures de rappel et un meilleur équilibre des métriques. Étant donné l'objectif de capter davantage d'opportunités associées à un état de réservation de classe minoritaire (prédictions positives), la réduction des faux négatifs est priorisée, ce qui justifie l'adoption de la méthode de pondération des données.

Tableau 5.3 Mesures de performance pour le modèle LightGBM par rapport à différentes régions en utilisant SMOTE

Région	Exactitude	Précision	Rappel	F1-score
Domestique	90%	60%	64%	62%
États-Unis	87%	69%	65%	67%
Atlantique	87%	76%	72%	74%
Sud	89%	80%	71%	75%
Pacifique	91%	78%	77%	77%

5.1.2 Analyse comparative des performances de différents modèles d'apprentissage automatique

Les résultats des différents modèles montrent des variations selon les régions et les métriques. LightGBM se distingue comme le modèle le plus performant dans la majorité des régions,

grâce à sa capacité à maintenir un équilibre entre les métriques. Dans la région Domestique, LightGBM obtient la meilleure exactitude de 92% et les mesures de performance les plus élevés, surpassant les autres modèles, comme indiqué dans le tableau 5.4. Les modèles k-NN et forêt aléatoire montrent des mesures de performance inférieures, tandis que la régression logistique et l'arbre de décision affichent des résultats faibles sur toutes les métriques.

Tableau 5.4 Comparaison des performances des modèles d'apprentissage automatique pour la région Domestique

Modèle	Exactitude	Précision	Rappel	F1-score
LightGBM	92%	64%	65%	64%
k-NN	90%	62%	59%	60%
Régression Logistique	82%	44%	38%	41%
Forêt Aléatoire	88%	58%	56%	57%
Arbre de Décision	85%	49%	45%	47%

Dans le tableau 5.5 relatif à la région des États-Unis, k-NN présente des mesures de performance proches de celles de LightGBM, mais avec un rappel inférieur de 65% contre 67%. Cependant, LightGBM conserve un avantage grâce à un F1-score de 68%. Les modèles moins performants, tels que la régression logistique, continuent de montrer des limites, particulièrement avec un faible F1-score de 47%, ce qui les rend peu fiables pour cette région.

Tableau 5.5 Comparaison des performances des modèles d'apprentissage automatique pour la région États-Unis

Modèle	Exactitude	Précision	Rappel	F1-score
LightGBM	87%	69%	67%	68%
k-NN	86%	67%	65%	66%
Régression Logistique	79%	49%	46%	47%
Forêt Aléatoire	86%	63%	60%	61%
Arbre de Décision	81%	55%	51%	53%

La région Pacifique montre, dans le tableau 5.6, que les deux modèles, LightGBM et k-NN, atteignent une exactitude de 93%. Cependant, LightGBM prend l'avantage avec un rappel de 81% et un F1-score de 79% légèrement supérieurs, ce qui en fait le modèle préféré. Les autres modèles, en particulier la régression logistique, présentent des performances limitées, avec un F1-score aussi bas que 50%.

Dans le tableau 5.7 relatif à la région Atlantique, k-NN se présente comme le modèle le plus performant avec une exactitude de 91% et le meilleur F1-score de 79%. Bien que LightGBM présente de bonnes mesures de performance, k-NN semble mieux adapté pour cette région. Les

Tableau 5.6 Comparaison des performances des modèles d'apprentissage automatique pour la région Pacifique

Modèle	Exactitude	Précision	Rappel	F1-score
LightGBM	93%	76%	81%	79%
k-NN	93%	75%	79%	77%
Régression Logistique	82%	47%	53%	50%
Forêt Aléatoire	90%	67%	76%	71%
Arbre de Décision	84%	52%	58%	55%

modèles comme l'arbre de décision et la régression logistique, bien qu'améliorés par rapport à d'autres régions, restent en retrait.

Tableau 5.7 Comparaison des performances des modèles d'apprentissage automatique pour la région Atlantique

Modèle	Exactitude	Précision	Rappel	F1-score
LightGBM	88%	74%	78%	76%
k-NN	91%	78%	80%	79%
Régression Logistique	75%	50%	51%	50%
Forêt Aléatoire	85%	68%	74%	70%
Arbre de Décision	77%	53%	61%	57%

Enfin, pour la région Sud présentée dans le tableau 5.8, LightGBM présente une exactitude de 91% et un F1-score de 77%, bien que k-NN reste proche avec des performances comparables. Les autres modèles, comme la forêt aléatoire et l'arbre de décision, sont dépassés dans cette région également, montrant des résultats moyens sur toutes les métriques.

Tableau 5.8 Comparaison des performances des modèles d'apprentissage automatique pour la région Sud

Modèle	Exactitude	Précision	Rappel	F1-score
LightGBM	91%	75%	79%	77%
k-NN	90%	73%	76%	74%
Régression Logistique	81%	49%	52%	51%
Forêt Aléatoire	88%	65%	71%	68%
Arbre de Décision	83%	54%	52%	53%

En conclusion, LightGBM s'affirme comme le modèle de choix grâce à sa performance stable et élevée dans la plupart des régions. Cependant, k-NN offre une alternative, dans la région Atlantique, où il surpasse LightGBM.

5.1.3 Comparaison des mesures de performance entre le modèle LightGBM et le modèle de référence

En comparant LightGBM et le modèle de référence en utilisant les tableaux 5.2 et 5.9, des différences émergent. Bien que le modèle de référence affiche une exactitude élevée dans toutes les régions, ses performances sont limitées en termes de précision, de rappel et de F1-score, avec des métriques inférieurs à ceux obtenus par LightGBM. Par exemple, LightGBM atteint un F1-score maximal de 79% dans la région Pacifique, contre seulement 44% pour le modèle de référence. Cela reflète une meilleure capacité de LightGBM à équilibrer la précision et le rappel, rendant ses prédictions plus fiables. En revanche, les faibles valeurs de rappel indique que le modèle de référence est biaisé par la classe majoritaire. Ainsi, LightGBM s'impose comme le choix préféré et justifie l'utilisation de la modélisation par apprentissage automatique.

Tableau 5.9 Mesures de performance pour le modèle de référence par rapport à différentes régions

Région	Exactitude	Précision	Rappel	F1-score
Domestique	88%	41%	8%	13%
États-Unis	89%	47%	13%	20%
Atlantique	93%	56%	19%	28%
Sud	95%	52%	26%	35%
Pacifique	97%	61%	34%	44%

5.2 Résultats de la deuxième stratégie de modélisation

Dans la région Pacifique, les performances des modèles présentés dans le tableau 5.10 sont élevées malgré l'impact des données déséquilibrées. Avec une exactitude moyenne de 93%, le modèle parvient à maintenir une prédiction globalement correcte. Cependant, la précision moyenne de 81% est inférieure au rappel moyen de 86%, ce qui reflète une tendance à prédire un plus grand nombre de cas dont l'état de réservation est égal à 1.

Étant donné que trois modèles ont été développés, chacun entraîné sur une partie de données spécifique basé sur la partition de la fenêtre de réservation, chaque modèle a appris un seuil optimal s différent qui maximise le F1-score. Les modèles entraînés sur des fenêtres de réservation plus larges présentent des valeurs de seuil plus élevées par rapport à ceux entraînés sur des fenêtres plus courtes, ce qui est intuitif. Pour une fenêtre de réservation de $241 \leq BW \leq 364$, le seuil s est de 0,8, ce qui est supérieur à sa valeur pour une fenêtre de réservation de $0 \leq BW \leq 120$, où il est égal à 0,45.

Tableau 5.10 Mesures de performance pour le modèle relatif à la région Pacifique en utilisant la deuxième stratégie

Fenêtre de réservation	Exactitude	Précision	Rappel	F1-score	s
$0 \leq BW \leq 120$	87%	83%	90%	86%	0.45
$121 \leq BW \leq 240$	93%	82%	86%	84%	0.625
$241 \leq BW \leq 364$	98%	78%	81%	79%	0.8
Moyenne	93%	81%	86%	83%	0.63

Pour la région Atlantique, l'exactitude moyenne de 92% est inférieure à celle du Pacifique, comme montré dans le tableau 5.11. La précision moyenne de 80% reste cohérente à travers les fenêtres de réservation, mais le rappel est plus modéré, avec une moyenne de 81%. Cette région montre une performance équilibrée, mais la précision plus faible pour $121 \leq BW \leq 240$ met en évidence l'impact du déséquilibre des classes, où le modèle présente des difficultés à éviter les faux positifs.

Tableau 5.11 Mesures de performance pour le modèle relatif à la région Atlantique en utilisant la deuxième stratégie

Fenêtre de réservation	Exactitude	Précision	Rappel	F1-score	s
$0 \leq BW \leq 120$	86%	81%	83%	82%	0.55
$121 \leq BW \leq 240$	93%	78%	80%	79%	0.675
$241 \leq BW \leq 364$	98%	81%	81%	81%	0.75
Moyenne	92%	80%	81%	81%	0.66

La région Sud atteint des résultats comparables au Pacifique, comme présenté dans le tableau 5.12, avec une exactitude moyenne de 93% et un F1-score de 82%. Ici, le rappel de 82% est supérieur à la précision 81%, reflétant un biais vers la prédiction de la classe minoritaire. Ce résultat est acceptable car il est préférable de favoriser la détection des cas minoritaires plutôt que de renforcer la prédominance de la classe majoritaire. Cependant, les fenêtres de réservation larges $241 \leq BW \leq 364$ montrent une diminution du F1-score de 81%, indiquant que l'impact des données déséquilibrées augmente à mesure que la fenêtre de réservation s'élargit. En termes d'évolution du seuil s pour les régions Sud et Atlantique, la conclusion reste la même que pour la région Pacifique, une fenêtre de réservation plus large correspondant à une valeur de seuil plus élevée.

La région États-Unis est affectée par les données déséquilibrées, avec une exactitude moyenne plus faible de 89% et un F1-score moyen de seulement 75%, comme montré dans le tableau 5.13. La précision de 75% et le rappel de 76% sont également les plus bas parmi toutes les régions, reflétant la difficulté du modèle à distinguer efficacement entre les classes. Ce

Tableau 5.12 Mesures de performance pour le modèle relatif à la région Sud en utilisant la deuxième stratégie

Fenêtre de réservation	Exactitude	Précision	Rappel	F1-score	s
$0 \leq BW \leq 120$	87%	82%	81%	82%	0.6
$121 \leq BW \leq 240$	94%	81%	84%	82%	0.7
$241 \leq BW \leq 364$	97%	80%	82%	81%	0.775
Moyenne	93%	81%	82%	82%	0.69

problème se trouve dans $241 \leq BW \leq 364$, où l'exactitude chute à 85%. Le seuil élevé $s = 0.9$ contribue à aggraver cette situation en rendant la classification des échantillons de la classe minoritaire plus difficile. Ce seuil tend toutefois à diminuer lorsque les fenêtres de réservation sont plus courtes, facilitant ainsi la détection de cette classe.

Tableau 5.13 Mesures de performance pour le modèle relatif à la région des États-Unis en utilisant la deuxième stratégie

Fenêtre de réservation	Exactitude	Précision	Rappel	F1-score	s
$0 \leq BW \leq 120$	88%	76%	76%	76%	0.65
$121 \leq BW \leq 240$	93%	76%	81%	78%	0.8
$241 \leq BW \leq 364$	85%	73%	72%	72%	0.9
Moyenne	89%	75%	76%	75%	0.78

La région Domestique présente une exactitude moyenne de 93%, mais sa précision moyenne est la plus faible parmi toutes les régions de 72%, comme indiqué dans le tableau 5.14. Ce résultat suggère une influence du déséquilibre de classes sur le comportement du modèle, qui tend à privilégier la prédiction de la classe majoritaire au détriment de la classe minoritaire. Cette tendance se traduit par un nombre élevé de faux positifs, où le modèle attribue incorrectement la classe majoritaire aux cas qui se trouvent dans la classe minoritaire. Cela se reflète dans un F1-score moyen de 73%. Bien que l'exactitude soit élevée, ces résultats montrent que les données déséquilibrées ont un impact sur la capacité du modèle à prédire la classe minoritaire de manière fiable. De plus, les régions Domestique et États-Unis présentent un seuil plus élevé pour différentes partitions de la fenêtre de réservation que les autres régions, atteignant 0.9 pour des fenêtres de réservation supérieures à 240 jours.

Les résultats montrent que les données déséquilibrées affectent les performances différemment selon les régions et les fenêtres de réservation. Les régions telles que le Pacifique et le Sud gèrent mieux le déséquilibre, en maintenant un bon équilibre entre la précision et le rappel. En revanche, les régions Domestique et États-Unis présentent certains difficultés, avec des mesures de précision et de F1-score plus faibles.

Tableau 5.14 Mesures de performance pour le modèle relatif à la région Domestique en utilisant la deuxième stratégie

Fenêtre de réservation	Exactitude	Précision	Rappel	F1-score	s
$0 \leq BW \leq 120$	85%	71%	77%	74%	0.6
$121 \leq BW \leq 240$	96%	72%	73%	72%	0.825
$241 \leq BW \leq 364$	99%	72%	76%	74%	0.9
Moyenne	93%	72%	75%	73%	0.78

Le tableau 5.15 montre les gains réalisés dans les mesures de performance de la deuxième stratégie par rapport à la première. La précision s'améliore peu par rapport aux autres mesures. La région Domestique et États-Unis sont les régions qui présentent la meilleure amélioration avec cette nouvelle stratégie.

En conclusion, les modèles issus de la deuxième stratégie présentent des mesures de performance supérieures à 70% pour l'ensemble des régions, ce qui les rend fiables pour les prochaines interprétations. Ces performances sont par ailleurs renforcées par l'application de la troisième stratégie.

Tableau 5.15 Gain en mesures de performance par la deuxième stratégie de modélisation

Région	Gain d'exactitude	Gain de la précision	Gain du rappel	Gain du F1-score
Domestique	1%	8%	10%	9%
États-Unis	2%	6%	9%	7%
Atlantique	4%	6%	3%	5%
Sud	2%	6%	3%	5%
Pacifique	0%	5%	5%	4%

5.2.1 Analyse du seuil de classification

La figure 5.1 illustre le processus d'optimisation du seuil de classification s dans le cadre de la validation croisée. Ce processus consiste à évaluer différentes valeurs possibles de s sur les probabilités estimées issues des plis de validation, afin d'identifier celle qui maximise le F1-score. Cette démarche a été appliquée au modèle associé à la région Domestique pour une fenêtre de réservation comprise entre 0 et 120 jours. Le seuil optimal identifié est $s = 0{,}6$, correspondant à un F1-score maximal de 74 %.

La courbe ROC est également présentée dans la figure 5.2 afin d'évaluer les performances du modèle à partir des valeurs de sensibilité et de spécificité associées au seuil de classification retenu. Cette courbe représente le taux de vrais positifs en axe des ordonnées, également

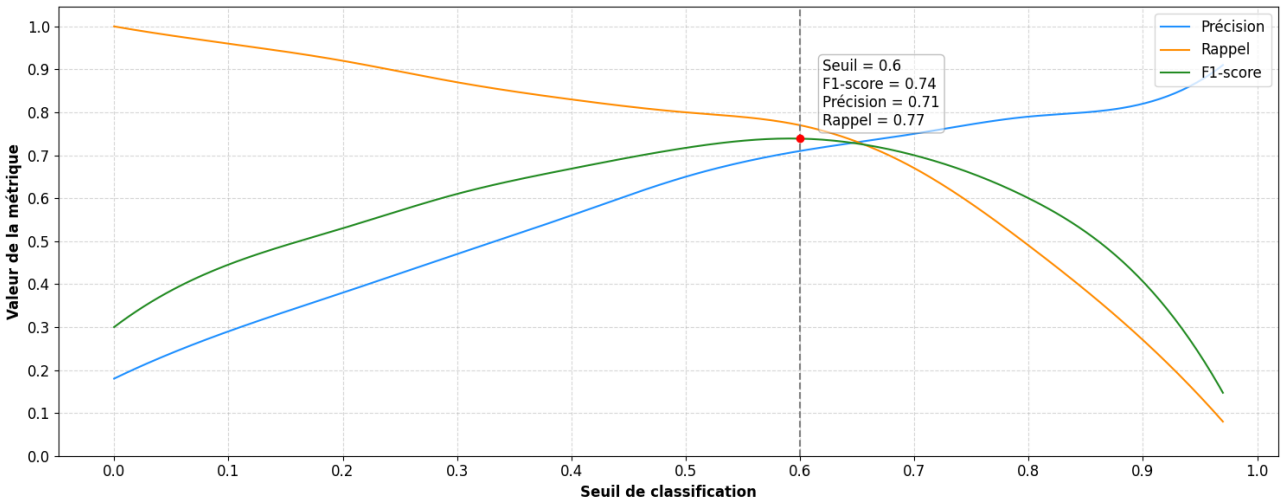


FIGURE 5.1 Optimisation du seuil de classification pour le modèle lié à la région Domestique et à la fenêtre de réservation entre 0 et 120 jours

appelé rappel ou sensibilité, en fonction du taux de faux positifs en axe des abscisses, lequel est défini comme 1 − spécificité. La spécificité est donnée par la formule suivante :

$$\text{Spécificité} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{5.1}$$

Chaque point de la courbe correspond à un seuil particulier utilisé pour convertir les probabilités estimées en prédictions. Dans cet exemple, le seuil optimal est égal à 0,6, ce qui correspond à une sensibilité de 77 % et une spécificité de 92 %. Ces résultats indiquent que le modèle est capable de détecter une grande majorité des cas positifs tout en limitant les fausses alertes, ce qui reflète un bon compromis entre rappel et rejet des faux positifs.

Nous analysons ensuite la variation du seuil optimal s selon les régions et les différentes partitions de la fenêtre de réservation utilisées pour l'entraînement des modèles. Pour chaque modèle, le seuil s correspond à la valeur qui maximise le F1-score dans la prédiction de l'état de réservation, conformément à la règle de décision définie dans l'équation 2.16. Par exemple, le modèle associé à la région Pacifique et à une fenêtre de réservation comprise entre 0 et 120 jours indique un seuil de classification égal à 0,6. La figure 5.3 illustre l'évolution de ce seuil en fonction des partitions de la fenêtre de réservation pour différents modèles. Les résultats mettent en évidence une relation inverse entre la proximité de la date de départ et la valeur optimale du seuil s. Plus précisément, pour des fenêtres de réservation courtes, entre 0 et 120 jours, le seuil moyen est de 0,57, ce qui permet de prédire une réservation de surplus à partir d'une probabilité relativement faible. En revanche, pour des fenêtres de réservation

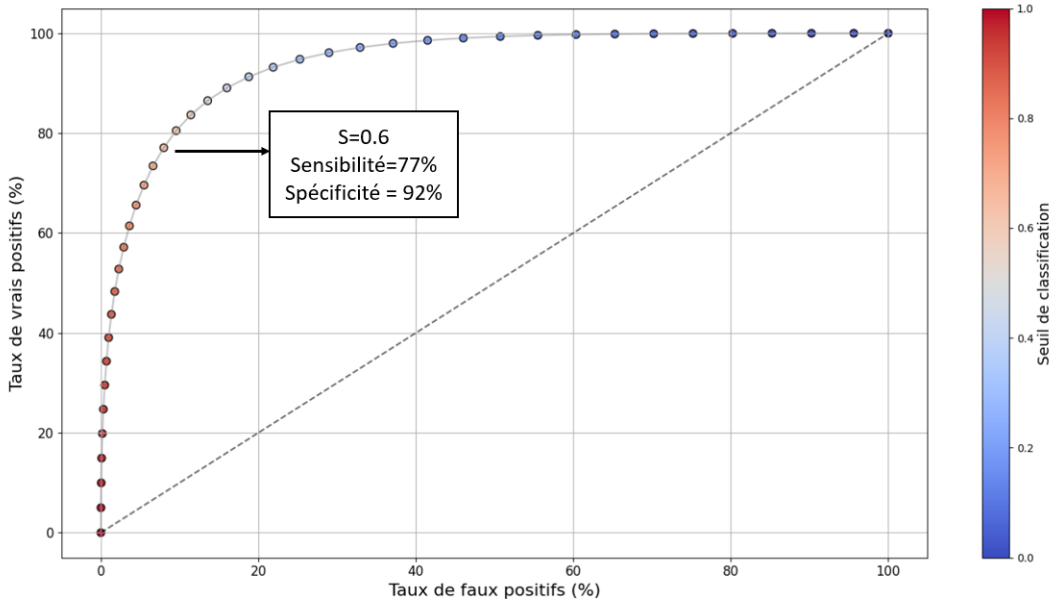


FIGURE 5.2 Courbe ROC pour le modèle lié à la région Domestique et à la fenêtre de réservation entre 0 et 120 jours

moyennes, entre 121 et 240 jours, ce seuil s'élève à 0,725, traduisant une exigence plus élevée en termes de probabilité pour classer un échantillon comme réservé.

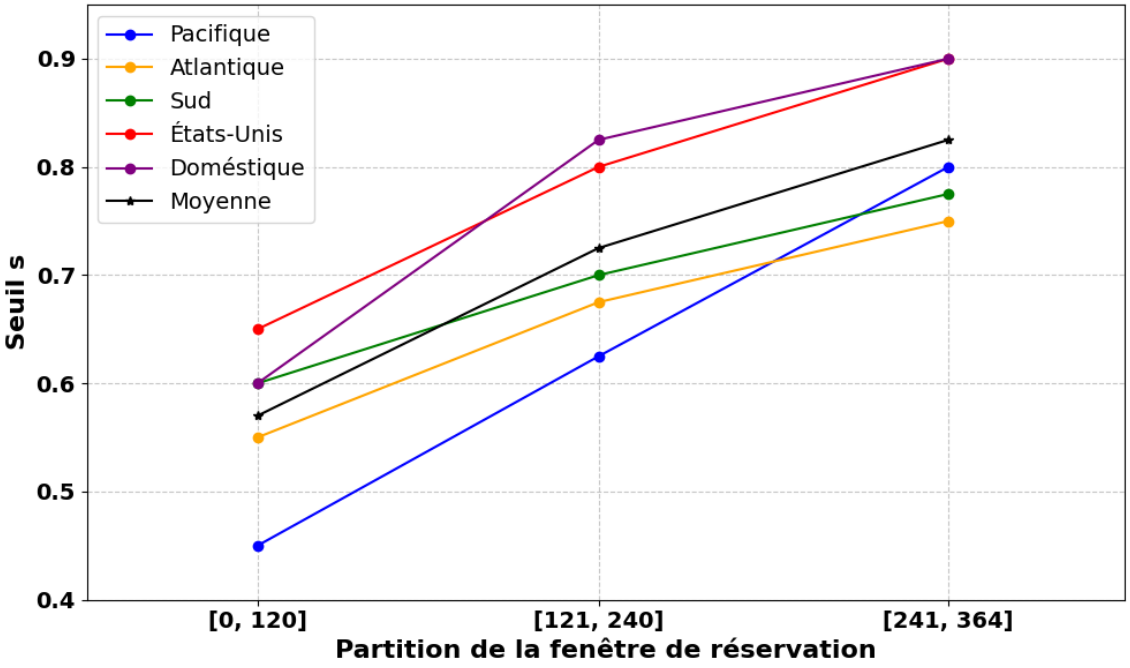


FIGURE 5.3 Analyse de l'évolution du seuil s relatif au F1-score optimal

À mesure que la date de départ s'éloigne, le seuil optimal continue d'augmenter, atteignant en moyenne 0,825 pour les fenêtres de réservation supérieures à 240 jours. Cette tendance s'explique par la raréfaction des réservations de surplus dans les longues fenêtres de réservation, ce qui nécessite un seuil de classification plus élevé pour confirmer une réservation. Inversement, lorsque la date de départ se rapproche, le seuil diminue, traduisant une tolérance accrue aux probabilités plus faibles.

Du point de vue régional, les zones Domestique et États-Unis présentent les seuils les plus élevés, toutes partitions confondues, indiquant qu'une probabilité plus importante est requise pour classer un échantillon comme classe minoritaire. À l'opposé, la région Pacifique affiche les seuils les plus bas, à l'exception des larges fenêtres de réservation, ce qui suggère qu'il est possible d'y observer des réservations de surplus même à des niveaux de probabilité plus faibles.

5.2.2 Évaluation du calibrage des probabilités

L'évaluation du calibrage des probabilités consiste à mesurer dans quelle mesure les probabilités estimées par le modèle sont alignées avec les probabilités empiriques observées. Cette analyse a été réalisée en utilisant l'ensemble du test. Avant d'aborder les méthodes d'évaluation, il convient de définir les deux notions clés suivantes :

- **Probabilités estimées** : correspondent aux probabilités générées par le modèle indiquant la probabilité qu'il existe au moins une réservation de surplus dans les sept prochains jours. Par exemple, si un modèle attribue une probabilité de 0,8 à un échantillon donné, cela signifie qu'il estime à 80 % la probabilité qu'il y ait une réservation de surplus.
- **Probabilités réelles** : désignent les probabilités empiriques observées pour un groupe d'échantillons ayant des probabilités estimées similaires. Elles sont calculées comme le ratio entre le nombre d'échantillons appartenant à la classe minoritaire et le nombre total d'échantillons dans ce groupe. Par exemple, si 80 % des échantillons ayant une probabilité estimée d'environ 0,8 présentent effectivement une réservation de surplus, alors la probabilité réelle correspondante est également de 0,8.

Le calibrage des probabilités est un processus qui consiste à ajuster les probabilités estimées du modèle afin qu'elles reflètent mieux la fréquence réelle des événements, selon Gupta et Ramdas [87]. Le Platt scaling est une technique de calibrage des probabilités fondée sur la régression logistique, utilisée pour ajuster les log-odds finaux produits par le modèle à partir de l'ensemble de test afin d'obtenir des probabilités calibrées. Dans ce cas, plutôt que d'utiliser l'équation 2.15 pour estimer directement la probabilité, l'équation suivante est utilisée à la

place :

$$Pr(y_i = 1 \mid x_i) = \frac{1}{1 + \exp(b_1 z_i^{(M)} + b_2)}, \tag{5.2}$$

où :

— $z_i^{(M)}$ désigne le log-odds final estimé par le modèle pour le i-ème échantillon.
— b_1 et b_2 sont les coefficients optimisés lors de l'entraînement de la régression logistique sur les paires $(z_i^{(M)}, y_i)$ issues d'un ensemble de validation.

Afin d'évaluer la qualité du calibrage des probabilités estimées, la perte de calibrage est analysée. Selon Stehouwer et al. [88], cette perte peut être quantifiée à l'aide du score de Brier, défini par l'équation suivante :

$$\text{Score de Brier} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (p_i - y_i)^2 . \tag{5.3}$$

Celle-ci quantifie la précision des probabilités estimées dans les modèles. Elle mesure l'erreur quadratique moyenne entre les probabilités estimées et les résultats réels. Une valeur plus faible du score de Brier indique un meilleur calibrage.

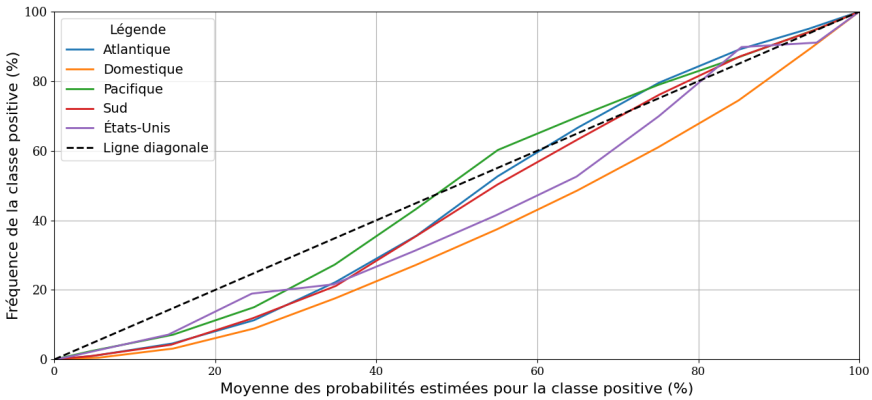
Par ailleurs, une courbe de calibrage est tracée pour visualiser cet alignement. Celle-ci représente, en abscisse, les probabilités estimées regroupées en intervalles (ou bacs), et en ordonnée, les probabilités empiriques observées pour chaque groupe. Un modèle parfaitement calibré générerait une courbe proche de la diagonale $y = x$, signifiant que les probabilités estimées correspondent aux fréquences réelles.

La figure 5.4 présente les courbes de calibrage obtenues sans application du Platt scaling, mais en utilisant les probabilités estimées par l'équation 2.15. Ces courbes se rapportent aux modèles pour différentes régions et partitions de la fenêtre de réservation.

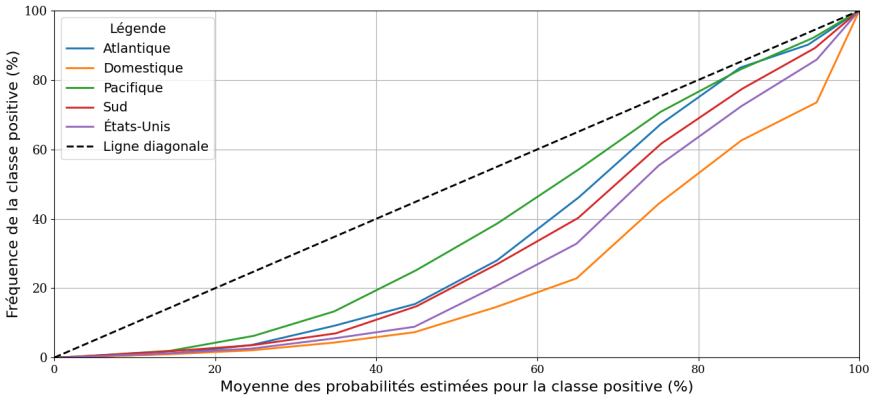
Il est observé que les courbes de calibrage tendent à se rapprocher davantage de la diagonale lorsque la fenêtre de réservation est courte. Cela suggère que les probabilités estimées deviennent plus représentatives des probabilités réelles à mesure que la date de départ approche.

Au niveau régional, les courbes associées aux régions États-Unis et Domestique présentent un éloignement plus marqué de la ligne diagonale, indiquant un calibrage initialement moins précis dans ces zones.

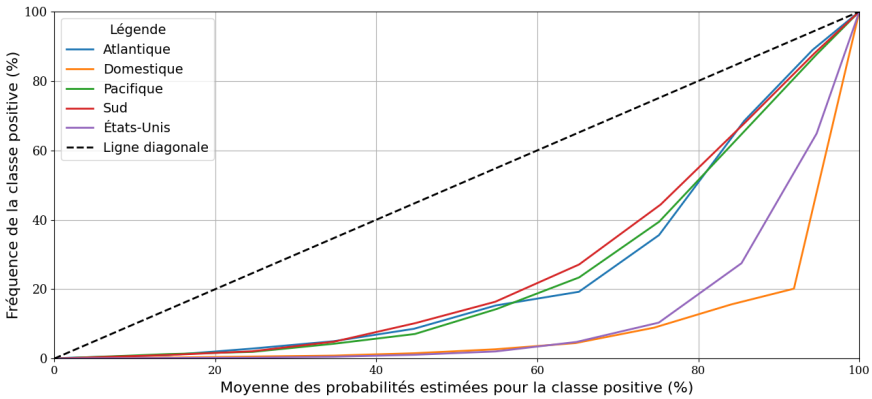
Après l'application de la méthode de Platt scaling, la figure 5.5 montre une amélioration notable : les courbes de calibrage sont plus proches de la diagonale, et ce pour l'ensemble



(a) Fenêtre de réservation entre 0 et 120 jours



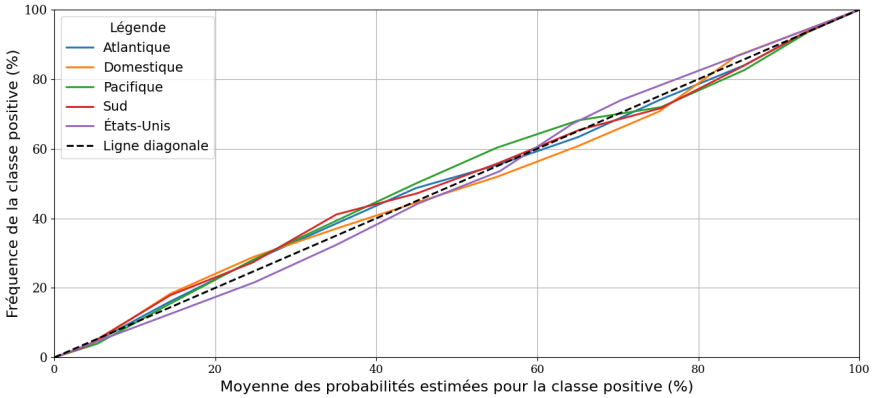
(b) Fenêtre de réservation entre 121 et 240 jours



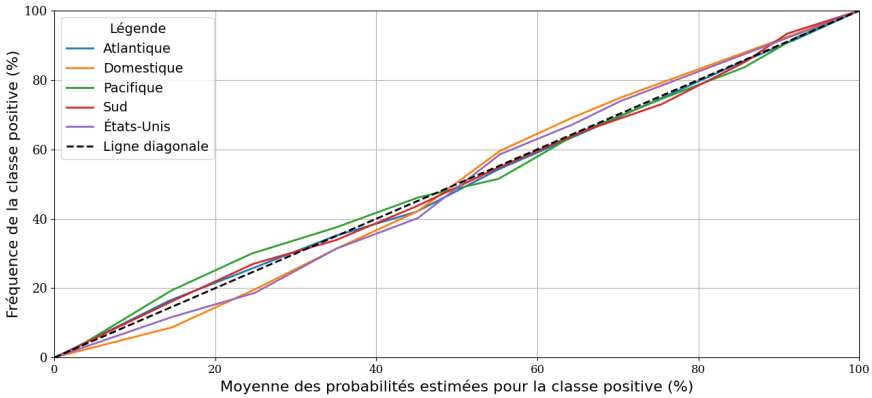
(c) Fenêtre de réservation entre 241 et 364 jours

FIGURE 5.4 Courbes de calibrage sans l'application de Platt scaling

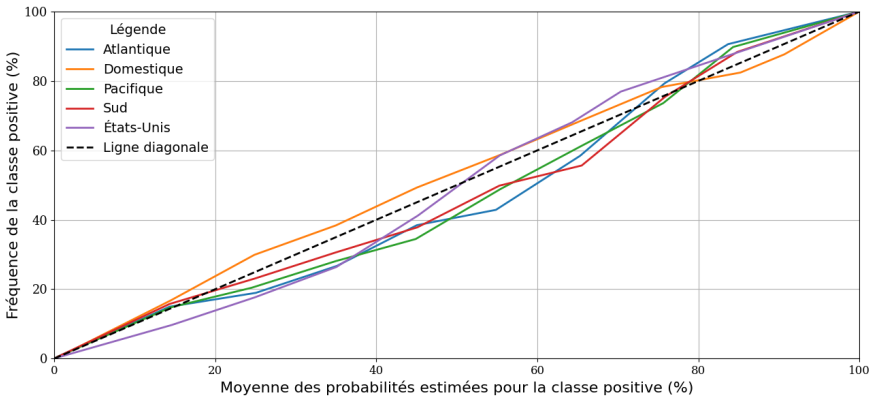
des partitions de la fenêtre de réservation et des régions. Cela confirme que les probabilités estimées reflètent mieux les fréquences empiriques observées.



(a) Fenêtre de réservation entre 0 et 120 jours



(b) Fenêtre de réservation entre 121 et 240 jours



(c) Fenêtre de réservation entre 241 et 364 jours

FIGURE 5.5 Courbes de calibrage avec l'application de Platt scaling

Concernant le score de Brier, la moyenne a été calculée pour chaque région, sans et avec l'application du Platt scaling, comme présenté dans le tableau 5.16. Les résultats montrent une réduction de ce score, ce qui témoigne d'une amélioration du calibrage des probabilités

estimées. Ainsi, l'application du Platt scaling permet de rapprocher les probabilités produites par le modèle des fréquences empiriques observées, renforçant la fiabilité des prédictions en sortie du modèle.

Tableau 5.16 Score de Brier sans et avec l'application de Platt scaling

Région	Sans l'application du Platt scaling	Avec l'application du Platt scaling
Domestique	0,210	0,0138
États-Unis	0,186	0,0092
Atlantique	0,116	0,0084
Sud	0,121	0,0071
Pacifique	0,096	0,0062

5.3 Résultats de la troisième stratégie de modélisation

Dans cette section, nous examinons les résultats de la troisième stratégie, dans laquelle les régions ont été regroupées en fonction du mois de départ et des partitions de la fenêtre de réservation. Pour simplifier la présentation des résultats et préparer la comparaison entre les stratégies, nous prenons, pour chaque région, la moyenne des mesures de performance par partition de la fenêtre de réservation.

Commençons par la région Domestique, on se concentre sur la comparaison entre la deuxième et de la troisième stratégie, à partir des tableaux 5.14 et 5.17. Le tableau 5.17 a été créé en prenant la moyenne des mesures de performance par partition de la fenêtre de réservation à partir des tableaux en annexe B.1, B.2 et B.3. La comparaison montre une amélioration des performances avec la troisième stratégie, notamment dans la réduction des faux positifs et l'augmentation des vrais positifs détectés. Dans toutes les partitions de la fenêtre de réservation, cette nouvelle stratégie affiche des précisions et rappels supérieurs, avec des gains particulièrement marqués pour $0 \leq BW \leq 120$ avec une précision de 71% à 81%, un rappel de 77% à 87%. En moyenne, le F1-score passe de 73% à 82%, reflétant un meilleur équilibre entre précision et rappel. Ces résultats indiquent que la troisième stratégie est plus efficace pour prédire l'état de réservation, en minimisant les erreurs et en améliorant la fiabilité des prédictions.

Pour la région États-Unis, on compare les tableaux 5.13 et 5.18 pour les deuxième et troisième stratégies respectivement. Le tableau 5.18 a été créé en prenant la moyenne des mesures de performance par partition de la fenêtre de réservation à partir des tableaux en annexe B.4, B.5 et B.6. La comparaison des deux stratégies révèle une amélioration des performances avec l'utilisation de la troisième stratégie. L'exactitude moyenne augmente de 89% à 96%,

Tableau 5.17 Mesures de performance pour le modèle basé sur la région Domestique en utilisant la troisième stratégie

Fenêtre de réservation	Exactitude	Précision	Rappel	F1-score	s
$0 \leq BW \leq 120$	91%	81%	87%	84%	0.6
$121 \leq BW \leq 240$	98%	81%	84%	82%	0.8
$241 \leq BW \leq 364$	99%	77%	81%	79%	0.9
Moyenne	96%	80%	84%	82%	0.78

ce qui reflète une meilleure gestion des vrais négatifs et des faux positifs, notamment pour la fenêtre $241 \leq BW \leq 364$ de 85% à 99%. La précision passe de 75% à 80%, indiquant une réduction des faux positifs, tandis que le rappel s'améliore de 76% à 84%, grâce à une meilleure identification des vrais positifs. Le F1-score, qui combine précision et rappel, progresse de 75% à 82%, soulignant l'équilibre supérieur entre les deux métriques. Ces résultats démontrent encore que la troisième stratégie est plus efficace.

Concernant l'évolution du seuil s dans les régions Domestique et États-Unis, les valeurs optimisées par les modèles dans la troisième stratégie sont proches de celles obtenues dans la deuxième stratégie. À mesure que la date de départ approche, ce seuil diminue.

Tableau 5.18 Mesures de performance pour le modèle basé sur la région États-Unis en utilisant la troisième stratégie

Fenêtre de réservation	Exactitude	Précision	Rappel	F1-score	s
$0 \leq BW \leq 120$	91%	79%	84%	81%	0.6
$121 \leq BW \leq 240$	98%	81%	84%	82%	0.8
$241 \leq BW \leq 364$	99%	80%	81%	80%	0.9
Moyenne	96%	80%	84%	82%	0.78

Ensuite, les tableaux 5.12 et 5.19 montrent la différence de résultats entre la deuxième et la troisième stratégie pour la région Sud. Le tableau 5.19 a été créé en prenant la moyenne des mesures de performance par partition de la fenêtre de réservation à partir des tableaux en annexe B.7, B.8 et B.9. La comparaison montre des améliorations par rapport au deuxiéme stratégie. L'exactitude moyenne passe de 93% à 95%, grâce à une meilleure gestion des vrais négatifs et des faux positifs, avec des gains marqués pour $0 \leq BW \leq 120$ de 87% à 91% et pour $121 \leq BW \leq 240$ de 94% à 98%. La précision augmente, de 81% à 83%, ce qui traduit une réduction des faux positifs. Le rappel passe de 82% à 87%, indiquant une meilleure détection des vrais positifs. Le F1-score, synthétisant précision et rappel, s'améliore de 82% à 85%. Ces résultats confirment aussi que la troisième stratégie améliore la qualité des prédictions pour l'état de réservation dans la région Sud. Le seuil s maintient la même

tendance évolutive à travers les partitions de la fenêtre de réservation, mais présente des valeurs plus faibles par rapport à celles observées dans les régions précédentes.

Tableau 5.19 Mesures de performance pour le modèle basé sur la région Sud en utilisant la troisième stratégie

Fenêtre de réservation	Exactitude	Précision	Rappel	F1-score	s
$0 \leq BW \leq 120$	91%	83%	89%	86%	0.5
$121 \leq BW \leq 240$	98%	84%	86%	85%	0.65
$241 \leq BW \leq 364$	97%	82%	85%	83%	0.7
Moyenne	95%	83%	87%	85%	0.6

En outre, les tableaux 5.10 et 5.20 montrent la différence de résultats entre la deuxième et la troisième stratégie pour la région Pacifique. Le tableau 5.20 a été créé en prenant la moyenne des mesures de performance par partition de la fenêtre de réservation à partir des tableaux en annexe B.10, B.11 et B.12. La troisième stratégie présente des performances supérieures ou similaires par rapport à la deuxième stratégie. L'exactitude reste stable en moyenne de 93%, avec une amélioration pour $0 \leq BW \leq 120$ de 87% à 89%. La précision moyenne passe de 81% à 83%, ce qui reflète une réduction des faux positifs, particulièrement pour $241 \leq BW \leq 364$ de 78% à 82%. Le rappel diminue en moyenne 86% à 85%, bien qu'il reste constant ou augmente pour certaines partitions de la fenêtre de réservation.

Le F1-score, qui équilibre précision et rappel, passe de 83% à 84% en moyenne, montrant une amélioration globale de la qualité des prédictions, avec une progression pour la fenêtre $241 \leq BW \leq 364$ de 79% à 82%. Ces résultats suggèrent également que la troisième stratégie améliore la précision tout en maintenant des performances globales cohérentes pour prédire l'état de réservation dans la région Pacifique.

En termes d'évolution du seuil s, les régions Pacifique et Atlantique présentent des valeurs proches à travers les partitions de la fenêtre de réservation. À mesure que la date de départ approche, ce seuil diminue.

Tableau 5.20 Mesures de performance pour le modèle basé sur la région Pacifique en utilisant la troisième stratégie

Fenêtre de réservation	Exactitude	Précision	Rappel	F1-score	s
$0 \leq BW \leq 120$	89%	86%	88%	87%	0.5
$121 \leq BW \leq 240$	92%	81%	87%	84%	0.57
$241 \leq BW \leq 364$	97%	82%	82%	82%	0.73
Moyenne	93%	83%	86%	84%	0.6

Enfin, la comparaison pour la région Atlantique est basée sur les tableaux 5.11 et 5.21 relati-

vement aux deux dernières stratégies. Le tableau 5.21 a été créé en prenant la moyenne des mesures de performance par partition de la fenêtre de réservation à partir des tableaux en annexe B.13, B.14 et B.15. La troisième stratégie améliore globalement les performances par rapport à la deuxième stratégie. L'exactitude moyenne augmente de 92% à 94%. La précision moyenne passe de 80% à 82%, indiquant une réduction des faux positifs. Le rappel moyen s'améliore également, passant de 81% à 85%, reflétant une meilleure détection des vrais positifs. Le F1-score, qui équilibre précision et rappel, augmente de 81% à 84%. Ces résultats montrent que la troisième stratégie offre une amélioration dans la prédiction de l'état de réservation pour la région Atlantique.

Tableau 5.21 Mesures de performance pour le modèle basé sur la région Atlantique en utilisant la troisième stratégie

Fenêtre de réservation	Exactitude	Précision	Rappel	F1-score	s
$0 \leq BW \leq 120$	90%	85%	90%	88%	0.5
$121 \leq BW \leq 240$	95%	82%	87%	85%	0.63
$241 \leq BW \leq 364$	97%	79%	79%	79%	0.73
Moyenne	94%	82%	85%	84%	0.6

Le tableau 5.22 montre le gain global des mesures de performance dans la troisième stratégie par rapport à la deuxième stratégie. On constate que le gain est le plus important dans la région Domestique et États-Unis, ce qui est important car ils ont les données les plus déséquilibrées et représentent 98 % des données. En revanche, le gain pour les autres régions est limité. En conclusion, avec ces derniers résultats de modélisation, de bonnes performances sont obtenues avec le moins d'erreurs possibles pour les différentes régions.

Tableau 5.22 Gain en mesures de performance par la troisième stratégie de modélisation

Région	Gain d'exactitude	Gain de la précision	Gain du rappel	Gain du F1-score
Domestique	3%	8%	9%	9%
États-Unis	7%	5%	8%	7%
Atlantique	2%	2%	4%	3%
Sud	0%	2%	3%	2%
Pacifique	0%	2%	0%	1%

5.4 Compréhension de l'effet des facteurs sur la probabilité de réserver de surplus dans les sept jours à venir

Dans cette section, l'effet de certains facteurs sur la probabilité de réservation de surplus dans les sept jours à venir est investigué.

En appliquant les méthodes mentionnées dans la méthodologie à des modèles de différentes régions relativement à la deuxième stratégie, étant donné qu'elle présente de bonnes mesures de performance, afin de faciliter la présentation des résultats.

5.4.1 Interprétation de l'importance des variables explicatives

L'importance d'une variable explicative reflète son influence sur la précision du modèle prédictif. Plus une variable explicative contribue à réduire l'impureté de Gini, plus elle est considérée comme importante. Dans cette recherche, nous analysons la contribution de diverses variables explicatives, telles que le mois de réservation, le jour de la semaine de départ et de réservation, l'heure de départ du vol et la distance parcourue, dans la prédiction de l'état de réservation. L'importance des variables explicatives est représentée sous forme de barres, où chaque variable est associée à un score qui indique sa contribution totale à la réduction de l'impureté de Gini à travers l'ensemble des arbres du modèle.

En appliquant cette méthode à des modèles de différentes régions relativement à la deuxième stratégie, nous obtenons les résultats présentés dans les figures en annexe C.1, C.2, C.3, C.4 et C.5.

La moyenne de l'importance de chaque variable explicative est prise pour les différentes régions. Les résultats sont présentés dans la figure 5.6. Il apparaît que la fenêtre de réservation est la variable la plus contributive, avec une contribution à la réduction totale de l'impureté de l'ordre de 10% en moyenne pour l'ensemble des modèles, soulignant l'importance de la saisonnalité dans les comportements de réservation. La deuxième variable la plus influente est la distance du vol, avec une contribution moyenne supérieure à 7%.

Les variables créées, telles que le nombre de jours avec d'allocations et de réservations, la partition de la vitesse de réservation moyenne, ainsi que le nombre maximal d'allocations, contribuent également au pouvoir prédictif du modèle. Leur contribution dépasse en moyenne 5%, ce qui les place parmi les variables les plus importantes du modèle.

D'autres variables, comme les mois de réservation et de départ, l'heure de départ, ainsi que les jours de la semaine associés à la réservation et au départ, montrent aussi une certaine contribution à la réduction de l'impureté. Toutefois, leur impact reste inférieur à celui des variables précédemment mentionnées. Enfin, des variables telles que le nombre final d'allocations, le taux final de réservation, le nombre final de réservations et l'itinéraire présentent des contributions moindres au pouvoir prédictif du modèle.

Les variables de la qualité de vol, extraites des vols de 2022, se révèlent utiles pour améliorer la précision des modèles prédictifs sur l'état de réservation.

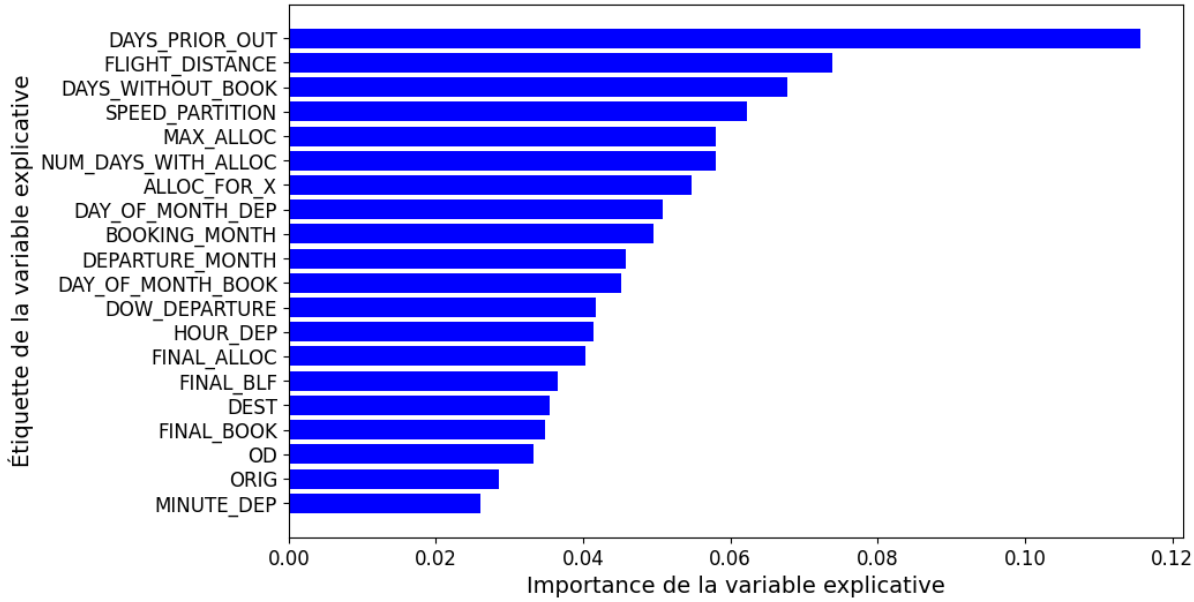


FIGURE 5.6 Importance moyenne des variables explicatives pour tous les modèles

5.4.2 Dépendance partielle des variables explicatives

L'analyse de dépendance partielle est réalisée séparément pour chaque région, en utilisant des modèles de la deuxième stratégie relatifs à une fenêtre de réservation comprise entre 0 et 120 jours, afin de capturer les différences régionales dans les comportements de réservation.

La figure 5.7 présente la dépendance partielle en fonction de la fenêtre de réservation. Une tendance similaire se dégage dans toutes les régions : plus la date de départ est proche, plus l'effet positif sur la probabilité de réservation de surplus est important, à l'exception des derniers jours, où un effet négatif est observé. Les valeurs les plus élevées de dépendance partielle sont observées entre 10 et 40 jours avant le départ, ce qui suggère que ces fenêtres de réservation ont l'influence la plus forte sur la probabilité de réservation de surplus. À partir de la fenêtre de réservation de 70 jours, plus la date de départ est éloignée, plus l'effet négatif, diminuant la probabilité estimée, est important.

De plus, la figure 5.8 présente la dépendance partielle en fonction de l'heure de départ. Dans toutes les régions, un effet négatif est observé durant les premières heures de la journée, cet effet étant particulièrement marqué dans les régions Sud, Domestique et États-Unis, où la dépendance partielle atteint des valeurs inférieures à $-0{,}1$. Entre 8 h et 20 h, il existe un effet positif sur la probabilité de réservation de surplus. Toutefois, cet effet ne se prolonge que jusqu'à 18 h pour la région des États-Unis, tandis qu'il persiste jusqu'à 22 h pour la région Pacifique. Enfin, durant les heures tardives, après 20 h, un nouvel effet négatif sur la

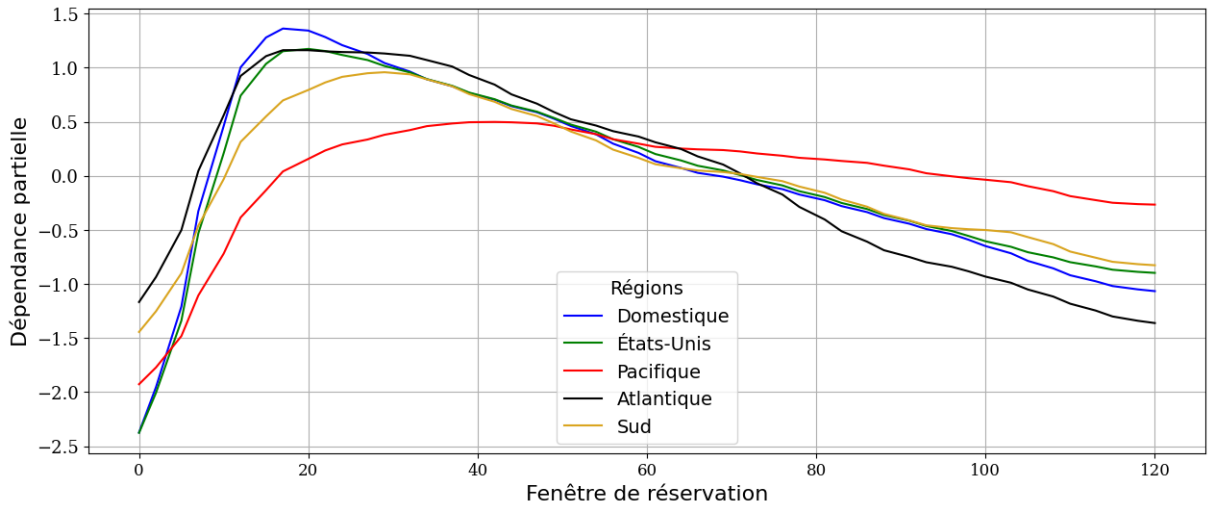


FIGURE 5.7 Dépendance partielle de la fenêtre de réservation

probabilité estimée est observé dans les régions Sud, Atlantique, Domestique et États-Unis.

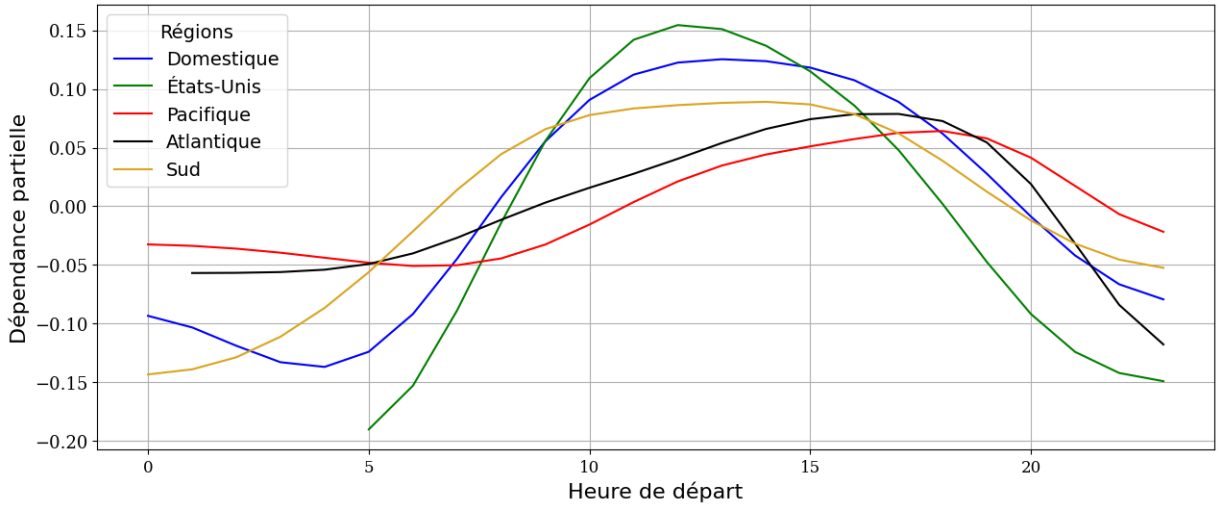


FIGURE 5.8 Dépendance partielle de l'heure de départ

Dans la figure 5.9, la dépendance partielle varie selon les différents mois de réservation, reflétant l'influence de cette variable sur la probabilité de réservation de surplus. Un effet positif est observé en janvier, février et mars pour l'ensemble des régions, à l'exception de la région Domestique. La région Sud présente un effet négatif d'avril à août. En fin d'année, un effet positif, augmentant la probabilité de réservation de surplus, est constaté dans les régions Pacifique, Sud et Domestique.

Enfin, la figure 5.10 illustre la dépendance partielle du jour de la semaine de la réservation.

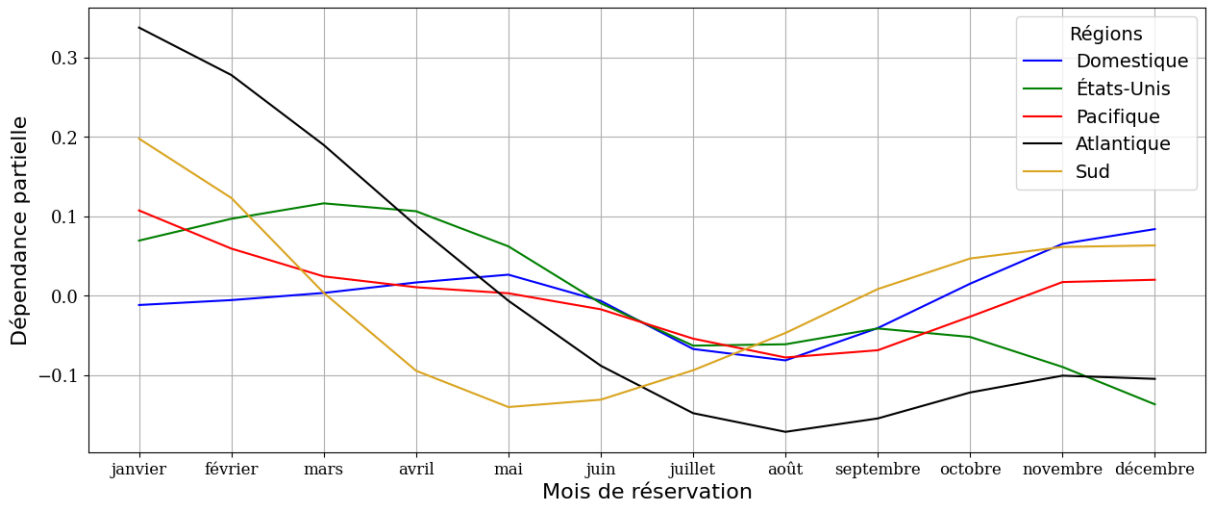


FIGURE 5.9 Dépendance partielle du mois de réservation

Dans l'ensemble des régions, les valeurs de dépendance partielle varient entre −0,025 et 0,15, ce qui indique une faible influence sur la probabilité de réservation de surplus. Un effet positif, augmentant la probabilité estimée, est observé en milieu de semaine, notamment les mercredis et jeudis.

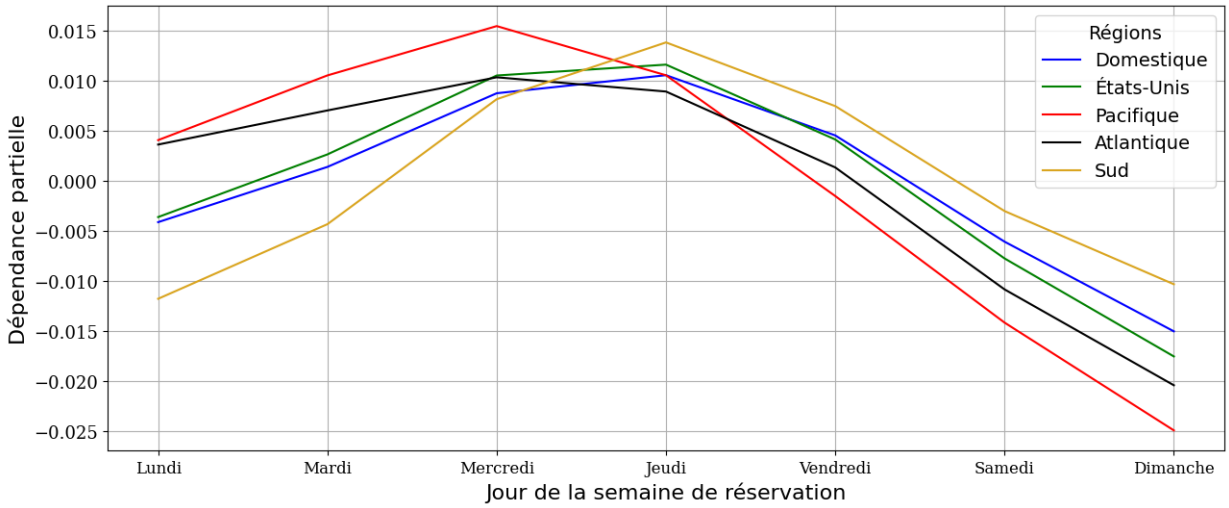


FIGURE 5.10 Dépendance partielle du jour de la semaine de réservation

5.5 Conclusion

Dans ce projet, la probabilité de réservation de sièges de surplus, considérés comme particulièrement rentables pour la compagnie aérienne, a été estimée à l'aide de modèles de clas-

sification. Plusieurs algorithmes ont été testés, notamment la régression logistique, l'arbre de décision, la forêt aléatoire, les k-plus proches voisins ainsi que LightGBM. Ce dernier s'est distingué par ses performances supérieures, en particulier grâce à sa capacité à gérer les données déséquilibrées. Il a surpassé le modèle de référence sur l'ensemble des métriques d'évaluation, confirmant ainsi la pertinence de l'apprentissage automatique dans ce contexte.

Afin d'améliorer la performance prédictive, trois stratégies de segmentation des données ont été mises en œuvre. Celles-ci ont permis de constituer des sous-ensembles plus homogènes, notamment selon la fenêtre de réservation et le mois de départ. La troisième stratégie, combinant ces deux dimensions, a donné les meilleurs résultats, notamment en termes de rappel. Par ailleurs, l'intégration de variables liées à la qualité des vols a renforcé la capacité des modèles à détecter les cas de réservation, ce qui s'est traduit par une amélioration de la précision des prédictions. Pour traiter le déséquilibre des classes, une pondération des données a été appliquée, s'avérant plus efficace que la technique SMOTE.

L'analyse de l'importance des variables explicatives a révélé que la fenêtre de réservation, la distance du vol et le nombre d'allocations comptent parmi les facteurs les plus déterminants dans la capacité prédictive du modèle. D'autres variables liées à la qualité des vols, telles que la vitesse de réservation moyenne, le nombre de jours avec de nouvelles réservations et le nombre maximal d'allocations, ont également contribué à affiner les prédictions.

En complément, une analyse de dépendance partielle a été réalisée afin d'examiner l'effet de certains facteurs sur la probabilité de réservation, tels que le mois de réservation, l'heure de départ et la fenêtre de réservation.

Ce projet s'est concentré exclusivement sur les sièges de classe économique. Une piste de recherche future consisterait à étendre l'analyse aux classes affaires et économique premium, dont les données sont davantage déséquilibrées. Une limite de cette recherche réside dans l'usage de l'encodage des étiquettes pour les variables temporelles cycliques telles que le jour de la semaine, le mois ou la partie du jour. Ce type d'encodage introduit une fausse hiérarchie entre les catégories et ne reflète pas leur nature périodique. L'utilisation d'un encodage cyclique aurait permis de mieux modéliser ces relations et d'améliorer la qualité des prédictions. Il serait également important de réduire l'horizon de prédiction actuellement fixé à sept jours, afin d'augmenter la réactivité opérationnelle du programme de fidélisation. Une réduction de cette période impliquerait toutefois de maintenir, voire d'améliorer, les performances observées. Enfin, l'approfondissement de l'interprétabilité locale des prédictions, notamment à travers l'analyse des contributions individuelles des variables dans le cadre de LightGBM, constituerait une étape importante pour renforcer la transparence du modèle et soutenir la prise de décision.

RÉFÉRENCES

[1] M. Wever, "Designing frequent flyer programs effectively - a market-research- and interview-based study for the german aviation sector," *The International Journal of Business and Management*, vol. 16, n°. 3, p. 58, févr. 2021. [En ligne]. Disponible : https://www.ccsenet.org/journal/index.php/ijbm/article/download/0/0/44705/47237

[2] A. Y. Orhun, T. Guo et A. Hagemann, "Reaching for gold : Frequent-flyer status incentives and moral hazard," *Marketing Science*, vol. 41, n°. 3, p. 548–574, janv. 2022. [En ligne]. Disponible : https://doi.org/10.1287/mksc.2021.1341

[3] C. Lee-Anant, "The importance of frequent flyer programs : An in-depth analysis," *Journal of Aerospace Technology and Management*, vol. 14, janv. 2022. [En ligne]. Disponible : https://doi.org/10.1590/jatm.v14.1254

[4] F. Alavi Fard, M. Sy et D. Ivanov, "Optimal overbooking strategies in the airlines using dynamic programming approach in continuous time," août 2019. [En ligne]. Disponible : https://www.sciencedirect.com/science/article/abs/pii/S1366554519300250

[5] A. Bazargan, S. Karray et S. Zolfaghari, "Can restrictions on redemption timing boost profitability of loyalty programs in competitive environments ?" *Computational Management Science*, vol. 18, n°. 1, p. 99–124, janv. 2021. [En ligne]. Disponible : https://doi.org/10.1007/s10287-020-00383-4

[6] M. Wever, "Designing frequent flyer programs effectively - a market-research- and interview-based study for the german aviation sector," *International Journal of Business and Management*, vol. 16, n°. 3, p. 58, févr. 2021. [En ligne]. Disponible : https://doi.org/10.5539/ijbm.v16n3p58

[7] S.-H. Park, M.-Y. Kim, Y.-J. Kim et Y.-H. Park, "A deep learning approach to analyze airline customer propensities : The case of south korea," *Applied Sciences*, vol. 12, n°. 4, p. 1916, févr. 2022. [En ligne]. Disponible : https://doi.org/10.3390/app12041916

[8] S. Lohiya, S. Ananthaselvi, A. Upade et S. Pai, "Aviation industry's dynamic pricing model (revenue management system) using data science," dans *2022 10th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-22)*, 2022, p. 1–6.

[9] A. E. Ertugrul et R. Şahin, "Emsrtrc : relaxation of booking limits by total revenue control for expected marginal seat revenue," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, n°. 3, p. 2221–2231, déc. 2022. [En ligne]. Disponible : https://link.springer.com/content/pdf/10.1007/s12652-022-04480-x.pdf

[10] A. Guerrini, G. Ferri, S. Rocchi, M. Cirelli, V. Pina et A. Grieszmann, “Personalization @ scale in airlines : combining the power of rich customer data, experiential learning, and revenue management,” *Journal of Revenue and Pricing Management*, vol. 22, n°. 2, p. 171–180, janv. 2023. [En ligne]. Disponible : https://link.springer.com/content/pdf/10.1057/s41272-022-00404-8.pdf

[11] E. Bachmat, S. Erland, F. Jaehn et S. Neumann, “Air passenger preferences : An international comparison affects boarding theory,” *Operations Research*, vol. 71, n°. 3, p. 798–820, nov. 2021. [En ligne]. Disponible : https://doi.org/10.1287/opre.2021.2148

[12] Z. Wang, X. Han, Y. Chen, X. Ye, K. Hu et D. Yu, “Prediction of willingness to pay for airline seat selection based on improved ensemble learning,” *Aerospace*, vol. 9, n°. 2, p. 47, janv. 2022. [En ligne]. Disponible : https://doi.org/10.3390/aerospace9020047

[13] X. Li, S. Gao, W. Yang, Y. Si et Z. Liu, “Purchase preferences-based air passenger choice behavior analysis from sales transaction data,” *Theoretical Computer Science*, vol. 928, p. 61–70, juin 2022. [En ligne]. Disponible : https://doi.org/10.1016/j.tcs.2022.06.013

[14] M. Pradana, A. Silvianita, P. N. Madiawati, D. Calandra, F. Lanzalonga et M. Oppioli, “A guidance to systematic literature review to young researchers by telkom university and the university of turin,” *To Maega | Jurnal Pengabdian Masyarakat*, vol. 6, n°. 2, p. 409, mai 2023. [En ligne]. Disponible : https://doi.org/10.35914/tomaega.v6i2.1915

[15] S. Kraus *et al.*, “Literature reviews as independent studies : guidelines for academic practice,” *Review of Managerial Science*, vol. 16, n°. 8, p. 2577–2595, oct. 2022. [En ligne]. Disponible : https://doi.org/10.1007/s11846-022-00588-8

[16] P. C. Sauer et S. Seuring, “How to conduct systematic literature reviews in management research : a guide in 6 steps and 14 decisions,” *Review of Managerial Science*, vol. 17, n°. 5, p. 1899–1933, mai 2023. [En ligne]. Disponible : https://doi.org/10.1007/s11846-023-00668-3

[17] M. Phillips, A. Van Epps, N. Johnson et D. Zwicky, “Effective engineering information literacy instruction : A systematic literature review,” *The Journal of Academic Librarianship*, vol. 44, n°. 6, p. 705–711, oct. 2018. [En ligne]. Disponible : https://doi.org/10.1016/j.acalib.2018.10.006

[18] X. Wang, H. Edison, D. Khanna et U. Rafiq, “How many papers should you review ? a research synthesis of systematic literature reviews in software engineering,” *arXiv (Cornell University)*, janv. 2023. [En ligne]. Disponible : https://arxiv.org/abs/2307.06056

[19] S. Kim et R. Giachetti, “A stochastic mathematical appointment overbooking model for healthcare providers to improve profits,” *IEEE Transactions on Systems Man and*

Cybernetics - Part a Systems and Humans, vol. 36, nº. 6, p. 1211–1219, oct. 2006. [En ligne]. Disponible : https://doi.org/10.1109/tsmca.2006.878970

[20] D. Carreras-García, D. Delgado-Gómez, E. Baca-García et A. Artés-Rodriguez, "A probabilistic patient scheduling model with time variable slots," *Computational and Mathematical Methods in Medicine*, vol. 2020, p. 1–10, sept. 2020. [En ligne]. Disponible : https://doi.org/10.1155/2020/9727096

[21] N. Sulima, "Probabilistic model of overbooking for an airline," *Automatic Control and Computer Sciences*, vol. 46, nº. 1, p. 49–56, févr. 2012. [En ligne]. Disponible : https://doi.org/10.3103/s0146411612010075

[22] S. O. Travin, Y. I. Skurlatov et A. V. Roshchin, "Capabilities and limitations of mathematical models in ecological safety forecasting," *Russian Journal of Physical Chemistry B*, vol. 14, nº. 1, p. 86–99, janv. 2020. [En ligne]. Disponible : https://doi.org/10.1134/s1990793120010315

[23] S. Shao, G. Kauermann et M. S. Smith, "Whether, when and which : Modelling advanced seat reservations by airline passengers," *Transportation Research Part a Policy and Practice*, vol. 132, p. 490–514, déc. 2019. [En ligne]. Disponible : https://doi.org/10.1016/j.tra.2019.12.002

[24] E. Chiew, R. A. Daziano et L. A. Garrow, "Bayesian estimation of hazard models of airline passengers' cancellation behavior," *Transportation Research Part a Policy and Practice*, vol. 96, p. 154–167, janv. 2017. [En ligne]. Disponible : https://doi.org/10.1016/j.tra.2016.12.006

[25] Y. Xie, "Values and limitations of statistical models," *Research in Social Stratification and Mobility*, vol. 29, nº. 3, p. 343–349, mai 2011. [En ligne]. Disponible : https://doi.org/10.1016/j.rssm.2011.04.001

[26] S. S. Henley, R. M. Golden et T. M. Kashner, "Statistical modeling methods : challenges and strategies," *Biostatistics Epidemiology*, vol. 4, nº. 1, p. 105–139, juill. 2019. [En ligne]. Disponible : https://doi.org/10.1080/24709360.2019.1618653

[27] D. B. Percival, "The spectral analysis of time series," *Technometrics*, vol. 40, nº. 4, p. 354, nov. 1998. [En ligne]. Disponible : https://doi.org/10.1080/00401706.1998.10485571

[28] J. Müller et K. Bogenberger, "Time series analysis of booking data of a free-floating carsharing system in berlin," *Transportation Research Procedia*, vol. 10, p. 345–354, janv. 2015. [En ligne]. Disponible : https://doi.org/10.1016/j.trpro.2015.09.084

[29] T.-H. Tsai et S. E. Kimes, *A Time Series Case-Based Predicting Model for Reservation Forecasting*, janv. 2009, p. 53–58. [En ligne]. Disponible : https://doi.org/10.1007/978-3-540-92814-0_9

[30] S. Jung, K.-M. Kim, H. Kwak et Y.-J. Park, “A worrying analysis of probabilistic time-series models for sales forecasting,” *arXiv (Cornell University)*, janv. 2020. [En ligne]. Disponible : https://arxiv.org/abs/2011.10715

[31] Z. Liu, Z. Zhu, J. Gao et C. Xu, “Forecast methods for time series data : a survey,” *IEEE Access*, vol. 9, p. 91896–91912, janv. 2021. [En ligne]. Disponible : https://doi.org/10.1109/access.2021.3091162

[32] S. O. A. Shah, “Optimizing hotel booking prediction : A comparative study of five machine learning algorithms,” *International Journal of Trendy Research in Engineering and Technology*, vol. 08, nº. 04, p. 31–41, janv. 2024. [En ligne]. Disponible : https://doi.org/10.54473/ijtret.2024.8406

[33] M. V. Rakesh, S. P. Kumar, N. Yogitha et R. Aishwarya, “Hotel booking cancelation prediction using ml algorithms,” *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, vol. 1, p. 466–471, févr. 2022. [En ligne]. Disponible : https://doi.org/10.1109/icais53314.2022.9742843

[34] M. A. Afrianto et M. Wasesa, “Booking prediction models for peer-to-peer accommodation listings using logistics regression, decision tree, k-nearest neighbor, and random forest classifiers,” *Journal of Information Systems Engineering and Business Intelligence*, vol. 6, nº. 2, p. 123, oct. 2020. [En ligne]. Disponible : https://doi.org/10.20473/jisebi.6.2.123-132

[35] P. Kumar et S. Sharma, “Hotel booking prediction using machine learning,” *International Journal for Research in Applied Science and Engineering Technology*, vol. 10, nº. 5, p. 4058–4060, mai 2022. [En ligne]. Disponible : https://doi.org/10.22214/ijraset.2022.43036

[36] W. Juntong et Y. Yin, “Hotel reservation status prediction based on hierarchical ensemble learning,” 2023. [En ligne]. Disponible : https://jglobal.jst.go.jp/en/detail?JGLOBAL_ID=202402240857322797

[37] M. A. Jishan, V. Singh, A. K. Ghosh, M. S. Alam, K. R. Mahmud et B. Paul, “Hotel booking cancellation prediction using applied bayesian models,” *arXiv (Cornell University)*, oct. 2024. [En ligne]. Disponible : http://arxiv.org/abs/2410.16406

[38] Z. Tang, “Prediction of the status of the hotel reservations,” *Applied and Computational Engineering*, vol. 8, nº. 1, p. 734–743, août 2023. [En ligne]. Disponible : https://doi.org/10.54254/2755-2721/8/20230154

[39] J. Zhang, D. Li, H. Lan, L. Tang et M. Guo, “Research on hotel reservation scheme based on random forest model prediction,” *Advances in Computer and Communication*, vol. 4, nº. 6, p. 358–362, janv. 2024. [En ligne]. Disponible : https://doi.org/10.26855/acc.2023.12.003

[40] Y. Chen, C. Ding, H. Ye et Y. Zhou, “Comparison and analysis of machine learning models to predict hotel booking cancellation,” *Advances in Economics, Business and Management Research/Advances in Economics, Business and Management Research*, janv. 2022. [En ligne]. Disponible : https://doi.org/10.2991/aebmr.k.220307.225

[41] H. Sun, Z. Lv, J. Li, Z. Xu et Z. Sheng, “Will the order be canceled ? order cancellation probability prediction based on deep residual model,” *Transportation Research Record Journal of the Transportation Research Board*, vol. 2677, n°. 6, p. 142–160, févr. 2023. [En ligne]. Disponible : https://doi.org/10.1177/03611981221144279

[42] D. Hopman, G. Koole et R. Van Der Mei, “A machine learning approach to itinerary-level booking prediction in competitive airline markets,” *International Journal of Revenue Management*, vol. 12, n°. 3/4, p. 153, janv. 2021. [En ligne]. Disponible : https://doi.org/10.1504/ijrm.2021.120347

[43] Y. Chen, C. Ding, H. Ye et Y. Zhou, “Comparison and analysis of machine learning models to predict hotel booking cancellation,” *Advances in Economics, Business and Management Research/Advances in Economics, Business and Management Research*, janv. 2022. [En ligne]. Disponible : https://doi.org/10.2991/aebmr.k.220307.225

[44] N. H. Zhu et N. W. Chen, “Comments on “a new ml based interference cancellation technique for layered space-time codes”,” *IEEE Transactions on Communications*, vol. 58, n°. 11, p. 3054–3055, sept. 2010. [En ligne]. Disponible : https://doi.org/10.1109/tcomm.2010.083110.100001

[45] P. B. Mpofu, G. Bakoyannis, C. T. Yiannoutsos, A. W. Mwangi et M. Mburu, “A pseudo-likelihood method for estimating misclassification probabilities in competing-risks settings when true-event data are partially observed,” *Biometrical Journal*, vol. 62, n°. 7, p. 1747–1768, juin 2020. [En ligne]. Disponible : https://doi.org/10.1002/bimj.201900198

[46] S. L. Harris et M. Samorani, “On selecting a probabilistic classifier for appointment no-show prediction,” *Decision Support Systems*, vol. 142, p. 113472, déc. 2020. [En ligne]. Disponible : https://doi.org/10.1016/j.dss.2020.113472

[47] J. H. Friedman, “Greedy function approximation : A gradient boosting machine.” *The Annals of Statistics*, vol. 29, n°. 5, oct. 2001. [En ligne]. Disponible : https://doi.org/10.1214/aos/1013203451

[48] Z. Huang et Z. Chen, “Comparison of different machine learning algorithms for predicting the sagd production performance,” *Journal of Petroleum Science and Engineering*, vol. 202, p. 108559, févr. 2021. [En ligne]. Disponible : https://doi.org/10.1016/j.petrol.2021.108559

[49] M. Osman, J. He, F. M. M. Mokbal, N. Zhu et S. Qureshi, “Ml-lgbm : A machine learning model based on light gradient boosting machine for the detection of version

number attacks in rpl-based networks," *IEEE Access*, vol. 9, p. 83654–83665, janv. 2021. [En ligne]. Disponible : https://doi.org/10.1109/access.2021.3087175

[50] H. Lu et R. Mazumder, "Randomized gradient boosting machine," *SIAM Journal on Optimization*, vol. 30, n°. 4, p. 2780–2808, janv. 2020. [En ligne]. Disponible : https://doi.org/10.1137/18m1223277

[51] M. S. Hosen et R. Amin, "Significant of gradient boosting algorithm in data management system," *Engineering International*, vol. 9, n°. 2, p. 85–100, juill. 2021. [En ligne]. Disponible : https://doi.org/10.18034/ei.v9i2.559

[52] Y. Qiu, J. Wang et Z. Li, "Personalized hrtf prediction based on lightgbm using anthropometric data," *China Communications*, vol. 20, n°. 6, p. 166–177, févr. 2023. [En ligne]. Disponible : https://doi.org/10.23919/jcc.2023.00.025

[53] D. A. McCarty, H. W. Kim et H. K. Lee, "Evaluation of light gradient boosted machine learning technique in large scale land use and land cover classification," *Environments*, vol. 7, n°. 10, p. 84, oct. 2020. [En ligne]. Disponible : https://doi.org/10.3390/environments7100084

[54] D. Kwak, Y. Liang, X. Shi et X. Tan, "Comparing machine learning and advanced methods with traditional methods to generate weights in inverse probability of treatment weighting : The inform study," *Pragmatic and Observational Research*, vol. Volume 15, p. 173–183, oct. 2024. [En ligne]. Disponible : https://doi.org/10.2147/por.s466505

[55] S. Hartini, Z. Rustam, G. S. Saragih et M. J. S. Vargas, "Estimating probability of banking crises using random forest," *IAES International Journal of Artificial Intelligence*, vol. 10, n°. 2, p. 407, mai 2021. [En ligne]. Disponible : https://doi.org/10.11591/ijai.v10.i2.pp407-413

[56] A. E. Maxwell, T. A. Warner et M. P. Strager, "Predicting palustrine wetland probability using random forest machine learning and digital elevation data-derived terrain variables," *Photogrammetric Engineering Remote Sensing*, vol. 82, n°. 6, p. 437–447, mai 2016. [En ligne]. Disponible : https://doi.org/10.14358/pers.82.6.437

[57] L. Fang et Y. Zhang, "Probability density function analysis based on logistic regression model," *International Journal of Circuits Systems and Signal Processing*, vol. 16, p. 60–69, janv. 2022. [En ligne]. Disponible : https://doi.org/10.46300/9106.2022.16.9

[58] N. Fuyane, M. Xaba et M. Sikwela, "Airline preference and choice factors in the south african domestic passenger market : An exploratory study," *The International Journal of Business Management*, vol. 13, n°. 1, p. 1–21, avr. 2021. [En ligne]. Disponible : https://sobiad.info/index.php/ijbms/article/view/472

[59] D. Banerji, V. Saha, N. Singh et R. Srivastava, "What are the most important consumer decision factors when choosing an airline ? an emerging economy perspective," *Asia*

Pacific Journal of Marketing and Logistics, vol. 35, n°. 1, p. 174–197, févr. 2022. [En ligne]. Disponible : https://doi.org/10.1108/apjml-07-2021-0486

[60] A. Kobaszyńska-Twardowska, M. Wantuła et A. Kinowski, "Impact of external factors on demand and supply in air transport in 2020-2021," *WUT Journal of Transportation Engineering*, vol. 134, p. 85–108, juin 2022. [En ligne]. Disponible : https://doi.org/10.5604/01.3001.0016.1443

[61] M. Aleksić, J. P. Raljić, T. Gajić, I. Blešić, M. Dragosavac, M. Penić et J. Bugarčić, "Factors of airline selection and reflight intention during the pandemic/case of serbian airlines users," *Frontiers in Psychology*, vol. 13, juill. 2022. [En ligne]. Disponible : https://doi.org/10.3389/fpsyg.2022.915321

[62] A. Ullah, S. R. Manzoor, S. Aziz et M. A. K. Niazi, "Factors affecting airlines e-ticket purchase intent in covid-19 pandemic," *Zenodo (CERN European Organization for Nuclear Research)*, janv. 2022. [En ligne]. Disponible : https://zenodo.org/record/5814640

[63] W. Fan, X. Wu, X. Y. Shi, C. Zhang, I. W. Hung, Y. K. Leung et L. S. Zeng, "Support vector regression model for flight demand forecasting," *International Journal of Engineering Business Management*, vol. 15, p. 184797902311743, janv. 2023. [En ligne]. Disponible : https://doi.org/10.1177/18479790231174318

[64] V. Lurkin, L. A. Garrow, M. J. Higgins, J. P. Newman et M. Schyns, "A comparison of departure time of day formulations," *SSRN Electronic Journal*, janv. 2016. [En ligne]. Disponible : https://doi.org/10.2139/ssrn.2729300

[65] G. Dutta et S. Santra, "An exploratory study of price movements along booking profiles in the airline industry in the indian domestic market," *International Journal of Revenue Management*, vol. 9, n°. 1, p. 72, janv. 2016. [En ligne]. Disponible : https://doi.org/10.1504/ijrm.2016.076186

[66] R. Brey et J. L. Walker, "Latent temporal preferences : An application to airline travel," *Transportation Research Part a Policy and Practice*, vol. 45, n°. 9, p. 880–895, juin 2011. [En ligne]. Disponible : https://doi.org/10.1016/j.tra.2011.04.010

[67] E. C. Chang, J. M. Pinegar et R. Ravichandran, "International evidence on the robustness of the day-of-the-week effect," *Journal of Financial and Quantitative Analysis*, vol. 28, n°. 4, p. 497, déc. 1993. [En ligne]. Disponible : https://doi.org/10.2307/2331162

[68] R. Sandhu et D. Klabjan, "Fleeting with passenger and cargo origin-destination booking control," *Transportation Science*, vol. 40, n°. 4, p. 517–528, nov. 2006. [En ligne]. Disponible : https://doi.org/10.1287/trsc.1060.0157

[69] S. Pölt, “How to aggregate origin and destination availability,” *Journal of Revenue and Pricing Management*, vol. 3, n°. 2, p. 191–199, juill. 2004. [En ligne]. Disponible : https://doi.org/10.1057/palgrave.rpm.5170106

[70] S. Birolini, M. Cattaneo, P. Malighetti et C. Morlotti, “Integrated origin-based demand modeling for air transportation,” *Transportation Research Part E Logistics and Transportation Review*, vol. 142, p. 102050, août 2020. [En ligne]. Disponible : https://doi.org/10.1016/j.tre.2020.102050

[71] Y. Yuan, L. Wu et X. Zhang, “Gini-impurity index analysis,” *IEEE Transactions on Information Forensics and Security*, vol. 16, p. 3154–3169, janv. 2021. [En ligne]. Disponible : https://doi.org/10.1109/tifs.2021.3076932

[72] R. Dunne, R. Reguant, P. Ramarao-Milne, P. Szul, L. M. Sng, M. Lundberg, N. A. Twine et D. C. Bauer, “Thresholding gini variable importance with a single-trained random forest : An empirical bayes approach,” *Computational and Structural Biotechnology Journal*, vol. 21, p. 4354–4360, janv. 2023. [En ligne]. Disponible : https://doi.org/10.1016/j.csbj.2023.08.033

[73] E. Scornet, “Trees, forests, and impurity-based variable importance in regression,” *Annales De L Institut Henri Poincaré Probabilités Et Statistiques*, vol. 59, n°. 1, janv. 2023. [En ligne]. Disponible : https://doi.org/10.1214/21-aihp1240

[74] J. H. Friedman et B. E. Popescu, “Predictive learning via rule ensembles,” *The Annals of Applied Statistics*, vol. 2, n°. 3, sept. 2008. [En ligne]. Disponible : https://doi.org/10.1214/07-aoas148

[75] T. Feizi, M. H. Moattar et H. Tabatabaee, “A multi-manifold learning based instance weighting and under-sampling for imbalanced data classification problems,” *Journal of Big Data*, vol. 10, n°. 1, oct. 2023. [En ligne]. Disponible : https://doi.org/10.1186/s40537-023-00832-2

[76] M. Han, A. Li, Z. Gao, D. Mu et S. Liu, “Hybrid sampling and dynamic weighting-based classification method for multi-class imbalanced data stream,” *Applied Sciences*, vol. 13, n°. 10, p. 5924, mai 2023. [En ligne]. Disponible : https://doi.org/10.3390/app13105924

[77] T. Nguyen, K. Mengersen, D. Sous et B. Liquet, “Smote-cd : Smote for compositional data,” *PLoS ONE*, vol. 18, n°. 6, p. e0287705, juin 2023. [En ligne]. Disponible : https://doi.org/10.1371/journal.pone.0287705

[78] V.-H. Truong, S. Tangaramvong et G. Papazafeiropoulos, “An efficient lightgbm-based differential evolution method for nonlinear inelastic truss optimization,” *Expert Systems With Applications*, vol. 237, p. 121530, sept. 2023. [En ligne]. Disponible : https://doi.org/10.1016/j.eswa.2023.121530

[79] T. Li, X. Hu, T. Li, J. Liao, L. Mei, H. Tian et J. Gu, “Enhanced prediction and evaluation of hydraulic concrete compressive strength using multiple soft computing and metaheuristic optimization algorithms,” *Buildings*, vol. 14, n°. 11, p. 3461, oct. 2024. [En ligne]. Disponible : https://doi.org/10.3390/buildings14113461

[80] R. P. Sheridan, A. Liaw et M. Tudor, “Light gradient boosting machine as a regression method for quantitative structure-activity relationships,” *arXiv (Cornell University)*, janv. 2021. [En ligne]. Disponible : https://arxiv.org/abs/2105.08626

[81] J. Zhang, D. Mucs, U. Norinder et F. Svensson, “Lightgbm : an effective and scalable algorithm for prediction of chemical toxicity–application to the tox21 and mutagenicity data sets,” *Journal of Chemical Information and Modeling*, vol. 59, n°. 10, p. 4150–4158, sept. 2019. [En ligne]. Disponible : https://doi.org/10.1021/acs.jcim.9b00633

[82] S. Li, N. Jin, A. Dogani, Y. Yang, M. Zhang et X. Gu, “Enhancing lightgbm for industrial fault warning : an innovative hybrid algorithm,” *Processes*, vol. 12, n°. 1, p. 221, janv. 2024. [En ligne]. Disponible : https://doi.org/10.3390/pr12010221

[83] Y. Wu, Y. Cheng, S. Guan, X. Wang, W. Shi, H. Xu, R. Lang, J. Xing, J. Zhu et Q. Chen, “Knn-based lead-free piezoelectric ceramics with high qm and enhanced d33 via a donor–acceptor codoping strategy,” *Inorganic Chemistry*, vol. 62, n°. 37, p. 15094–15103, sept. 2023. [En ligne]. Disponible : https://doi.org/10.1021/acs.inorgchem.3c02046

[84] Q. Ma, “Recent applications and perspectives of logistic regression modelling in healthcare,” *Theoretical and Natural Science*, vol. 36, n°. 1, p. 185–190, juill. 2024. [En ligne]. Disponible : https://doi.org/10.54254/2753-8818/36/20240614

[85] K. Kirasich, T. Smith et B. Sadler, “Random forest vs logistic regression : Binary classification for heterogeneous datasets,” *SMU Data Science Review*, vol. 1, n°. 3, p. 9, janv. 2018. [En ligne]. Disponible : https://scholar.smu.edu/cgi/viewcontent.cgi?article=1041&context=datasciencereview

[86] X. Huang, “Predictive models : Regression, decision trees, and clustering,” *Applied and Computational Engineering*, vol. 79, n°. 1, p. 124–133, juill. 2024. [En ligne]. Disponible : https://doi.org/10.54254/2755-2721/79/20241551

[87] C. Gupta et A. Ramdas, “Online platt scaling with calibeating,” *arXiv (Cornell University)*, janv. 2023. [En ligne]. Disponible : https://arxiv.org/abs/2305.00070

[88] N. Stehouwer, A. Rowland-Seymour, L. Gruppen, J. M. Albert et K. Qua, “Validity and reliability of brier scoring for assessment of probabilistic diagnostic reasoning,” *Diagnosis*, oct. 2024. [En ligne]. Disponible : https://doi.org/10.1515/dx-2023-0109

ANNEXE A DESCRIPTION DES SYMBOLES UTILISÉS

TABLEAU A.1 Description des symboles

Symbole	Description
D_{2023}	Ensemble de données pour les vols en 2023
D_{2022}	Ensemble de données pour les vols en 2022
D_f	Ensemble de données final
m	Nombre des variables explicatives, caractéristiques ou colonnes
n	Nombre des échantillons dans l'ensemble de données total
n_{train}	Nombre d'échantillons dans l'ensemble de données d'apprentissage
n_{test}	Nombre d'échantillons dans l'ensemble de données de test
x_j	Caractéristique, colonne ou variable explicative de l'indice j
x_i	Échantillon ou ligne de l'indice i
$x_{i,j}$	Donnée à l'intersection de la colonne j et de la ligne i
$\{x_j \mid j = 1, 2, \ldots, m\}$	Ensemble de variables explicatives
$\{x_i \mid i = 1, 2, \ldots, n\}$	Ensemble des échantillons
y	Variable cible : état de réservation dans les sept jours suivants
$\hat{y}$	Prédictions de l'état de réservation dans les sept jours suivants
x_r	Variable de réservation incrémentale
e	Indice d'échelonnement de fenêtre de réservation
p_i	Probabilité estimée de réservation de surplus pour i-ème échantillon
z_i	Logarithme des cotes (log-odds) de la probabilité de réservation de surplus
M	Nombre total d'apprenants faibles
A	Nombre total de classes de la variable cible
$\{g_1, g_2, \ldots, g_U\}$	Catégories distinctes pour une variable catégorielle
$l_h \in \{0, 1, 2, \ldots, U-1\}$	Etiquette numérique pour une catégorie g_h
$V_h = [0, 0, \ldots, 1, \ldots, 0]$	Représentation d'une catégorie c_h par un vecteur binaire de longueur u
$Corr_{j_1,j_2}$	Corrélation entre deux caractéristiques x_{j_1} et x_{j_2}
s	Seuil de probabilité correspond au F1-score le plus élevé
λ_i	Poids attribué à la i-ème échantillon en fonction de la distribution des classes
k	Nombre d'échantillons plus proches voisins en utilisant la distance euclidienne

ANNEXE B RÉSULTATS DE LA MODÉLISATION POUR LA TROISIÈME STRATÉGIE

TABLEAU B.1 Mesures de performance pour les modèles relatifs à la région Domestique avec une fenêtre de réservation entre 0 et 120 jours

Mois	1	2	3	4	5	6	7	8	9	10	11	12	Moyenne
Exactitude	91%	91%	91%	92%	91%	90%	91%	92%	91%	92%	92%	91%	91%
Précision	80%	81%	80%	82%	80%	80%	81%	81%	80%	81%	82%	80%	81%
Rappel	87%	89%	90%	87%	86%	87%	84%	86%	85%	87%	87%	89%	87%
F1-score	83%	85%	85%	84%	83%	83%	82%	83%	83%	83%	84%	85%	84%
s	0.55	0.575	0.55	0.6	0.6	0.625	0.6	0.6	0.6	0.6	0.575	0.575	0.6

TABLEAU B.2 Mesures de performance pour les modèles relatifs à la région Domestique avec une fenêtre de réservation entre 121 et 240 jours

Mois	1	2	3	4	5	6	7	8	9	10	11	12	Moyenne
Exactitude	99%	99%	99%	99%	99%	98%	97%	97%	98%	98%	98%	99%	98%
Précision	82%	81%	81%	80%	81%	81%	80%	79%	80%	80%	81%	81%	81%
Rappel	84%	85%	86%	83%	85%	82%	82%	80%	85%	83%	82%	86%	84%
F1-score	83%	83%	83%	82%	83%	82%	81%	80%	82%	82%	82%	83%	82%
s	0.8	0.825	0.8	0.8	0.8	0.825	0.75	0.75	0.775	0.8	0.825	0.8	0.8

Tableau B.3 Mesures de performance pour les modèles relatifs à la région Domestique avec une fenêtre de réservation entre 241 et 364 jours

Mois	1	2	3	4	5	6	7	8	9	10	11	12	Moyenne
Exactitude	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%
Précision	81%	79%	79%	78%	75%	68%	74%	74%	77%	77%	77%	81%	77%
Rappel	82%	82%	81%	78%	78%	75%	79%	86%	85%	80%	85%	81%	81%
F1-score	82%	81%	80%	78%	77%	71%	77%	80%	81%	78%	81%	81%	79%
s	0.9	0.875	0.9	0.9	0.925	0.9	0.9	0.85	0.85	0.875	0.875	0.9	0.9

Tableau B.4 Mesures de performance pour les modèles relatifs à la région États-Unis avec une fenêtre de réservation entre 0 et 120 jours

Mois	1	2	3	4	5	6	7	8	9	10	11	12	Moyenne
Exactitude	91%	91%	91%	91%	90%	91%	92%	92%	91%	90%	90%	90%	91%
Précision	80%	81%	80%	79%	78%	79%	77%	78%	77%	78%	81%	80%	79%
Rappel	85%	83%	82%	81%	82%	81%	83%	83%	86%	88%	88%	88%	84%
F1-score	82%	82%	81%	80%	80%	80%	80%	80%	81%	83%	84%	84%	81%
s	0.6	0.6	0.6	0.6	0.65	0.65	0.65	0.625	0.6	0.55	0.55	0.575	0.6

Tableau B.5 Mesures de performance pour les modèles relatifs à la région États-Unis avec une fenêtre de réservation entre 121 et 240 jours

Mois	1	2	3	4	5	6	7	8	9	10	11	12	Moyenne
Exactitude	99%	99%	99%	99%	99%	99%	98%	98%	98%	98%	99%	98%	98%
Précision	82%	80%	79%	82%	83%	77%	80%	80%	82%	80%	82%	82%	81%
Rappel	86%	86%	84%	81%	83%	85%	82%	83%	82%	81%	84%	85%	84%
F1-score	84%	83%	82%	81%	83%	81%	81%	82%	82%	80%	83%	83%	82%
s	0.8	0.825	0.8	0.85	0.8	0.75	0.75	0.75	0.8	0.8	0.85	0.75	0.8

Tableau B.6 Mesures de performance pour les modèles relatifs à la région États-Unis avec une fenêtre de réservation entre 241 et 364 jours

Mois	1	2	3	4	5	6	7	8	9	10	11	12	Moyenne
Exactitude	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%
Précision	84%	86%	72%	76%	74%	79%	80%	81%	81%	82%	82%	81%	80%
Rappel	81%	82%	79%	82%	79%	78%	79%	85%	81%	83%	86%	79%	81%
F1-score	82%	84%	75%	79%	76%	78%	80%	83%	81%	83%	84%	80%	80%
s	0.9	0.9	0.9	0.85	0.9	0.9	0.9	0.85	0.875	0.85	0.85	0.85	0.89

Tableau B.7 Mesures de performance pour les modèles relatifs à la région Sud avec une fenêtre de réservation entre 0 et 120 jours

Mois	1	2	3	4	5	6	7	8	9	10	11	12	Moyenne
Exactitude	89%	90%	89%	91%	91%	93%	89%	91%	94%	91%	91%	92%	91%
Précision	84%	78%	85%	84%	84%	85%	79%	83%	88%	84%	83%	88%	83%
Rappel	90%	91%	88%	92%	91%	88%	88%	89%	87%	88%	91%	89%	89%
F1-score	87%	89%	87%	88%	87%	78%	83%	86%	87%	86%	87%	88%	86%
s	0.45	0.5	0.5	0.5	0.45	0.55	0.5	0.55	0.6	0.55	0.55	0.5	0.5

Tableau B.8 Mesures de performance pour les modèles relatifs à la région Sud avec une fenêtre de réservation entre 121 et 240 jours

Mois	1	2	3	4	5	6	7	8	9	10	11	12	Moyenne
Exactitude	95%	96%	96%	96%	95%	96%	93%	94%	97%	95%	95%	96%	95%
Précision	85%	84%	82%	86%	83%	87%	87%	82%	85%	79%	84%	86%	84%
Rappel	85%	88%	86%	86%	93%	90%	86%	80%	89%	85%	88%	87%	86%
F1-score	85%	86%	84%	86%	88%	89%	87%	81%	87%	82%	86%	87%	85%
s	0.7	0.55	0.7	0.7	0.55	0.75	0.65	0.75	0.65	0.65	0.55	0.65	0.65

Tableau B.9 Mesures de performance pour les modèles relatifs à la région Sud avec une fenêtre de réservation entre 241 et 364 jours

Mois	1	2	3	4	5	6	7	8	9	10	11	12	Moyenne
Exactitude	97%	98%	98%	99%	99%	99%	99%	97%	97%	98%	95%	97%	97%
Précision	79%	84%	81%	81%	83%	83%	81%	86%	82%	84%	79%	82%	82%
Rappel	90%	87%	91%	86%	75%	83%	81%	85%	85%	89%	87%	91%	85%
F1-score	84%	85%	86%	83%	79%	83%	81%	86%	84%	86%	83%	86%	83%
s	0.55	0.65	0.55	0.65	0.85	0.9	0.8	0.8	0.75	0.7	0.65	0.55	0.7

Tableau B.10 Mesures de performance pour les modèles relatifs à la région Pacifique avec une fenêtre de réservation entre 0 et 120 jours

Mois	1	2	3	4	5	6	7	8	9	10	11	12	Moyenne
Exactitude	91%	87%	92%	90%	89%	91%	88%	90%	90%	91%	89%	88%	89%
Précision	91%	85%	89%	87%	85%	83%	80%	86%	87%	86%	84%	88%	86%
Rappel	93%	84%	85%	94%	91%	87%	74%	88%	92%	91%	89%	91%	88%
F1-score	92%	84%	87%	90%	88%	85%	77%	87%	89%	88%	86%	89%	87%
s	0.5	0.55	0.65	0.45	0.45	0.5	0.55	0.5	0.5	0.55	0.5	0.45	0.5

Tableau B.11 Mesures de performance pour les modèles relatifs à la région Pacifique avec une fenêtre de réservation entre 121 et 240 jours

Mois	1	2	3	4	5	6	7	8	9	10	11	12	Moyenne
Exactitude	91%	93%	93%	93%	92%	93%	92%	92%	90%	90%	91%	89%	92%
Précision	78%	77%	84%	81%	82%	85%	81%	83%	80%	87%	84%	80%	81%
Rappel	83%	83%	82%	88%	87%	90%	90%	92%	87%	81%	88%	88%	87%
F1-score	80%	80%	83%	85%	84%	88%	85%	87%	83%	84%	86%	84%	84%
s	0.65	0.65	0.65	0.6	0.55	0.55	0.55	0.45	0.45	0.65	0.55	0.55	0.57

Tableau B.12 Mesures de performance pour les modèles relatifs à la région Pacifique avec une fenêtre de réservation entre 241 et 364 jours

Mois	1	2	3	4	5	6	7	8	9	10	11	12	Moyenne
Exactitude	97%	98%	98%	99%	98%	98%	96%	97%	96%	95%	99%	98%	97%
Précision	80%	78%	82%	85%	82%	77%	83%	85%	85%	77%	83%	88%	82%
Rappel	83%	87%	83%	85%	82%	84%	78%	78%	80%	83%	81%	76%	82%
F1-score	82%	82%	83%	85%	82%	81%	80%	82%	83%	80%	82%	82%	82%
s	0.7	0.7	0.7	0.75	0.75	0.7	0.75	0.75	0.7	0.65	0.8	0.825	0.73

Tableau B.13 Mesures de performance pour les modèles relatifs à la région Atlantique avec une fenêtre de réservation entre 0 et 120 jours

Mois	1	2	3	4	5	6	7	8	9	10	11	12	Moyenne
Exactitude	91%	91%	90%	89%	88%	92%	89%	91%	89%	89%	90%	91%	90%
Précision	86%	85%	88%	86%	87%	83%	83%	88%	83%	83%	84%	87%	85%
Rappel	92%	93%	88%	92%	88%	89%	95%	93%	89%	89%	91%	90%	90%
F1-score	89%	89%	88%	89%	88%	85%	88%	90%	86%	86%	87%	88%	88%
s	0.5	0.45	0.55	0.45	0.5	0.5	0.45	0.45	0.5	0.55	0.45	0.5	0.5

Tableau B.14 Mesures de performance pour les modèles relatifs à la région Atlantique avec une fenêtre de réservation entre 121 et 240 jours

Mois	1	2	3	4	5	6	7	8	9	10	11	12	Moyenne
Exactitude	98%	98%	95%	95%	94%	99%	95%	94%	93%	94%	96%	95%	95%
Précision	89%	84%	77%	86%	81%	77%	87%	89%	80%	81%	80%	83%	82%
Rappel	82%	93%	87%	91%	88%	85%	87%	87%	87%	84%	86%	89%	87%
F1-score	86%	88%	82%	89%	84%	81%	87%	88%	83%	83%	83%	86%	85%
s	0.8	0.55	0.5	0.6	0.6	0.75	0.65	0.65	0.55	0.65	0.65	0.6	0.63

Tableau B.15 Mesures de performance pour les modèles relatifs à la région Atlantique avec une fenêtre de réservation entre 241 et 364 jours

Mois	1	2	3	4	5	6	7	8	9	10	11	12	Moyenne
Exactitude	98%	98%	98%	96%	98%	98%	97%	97%	98%	97%	98%	98%	97%
Précision	76%	75%	77%	70%	79%	81%	75%	84%	85%	82%	84%	83%	79%
Rappel	77%	79%	80%	65%	76%	90%	87%	76%	85%	85%	77%	75%	79%
F1-score	77%	77%	78%	67%	78%	85%	81%	80%	85%	83%	81%	79%	79%
s	0.75	0.7	0.65	0.85	0.75	0.75	0.6	0.75	0.75	0.7	0.8	0.825	0.73

ANNEXE C IMPORTANCE DES VARIABLES EXPLICATIVES POUR LES DIFFÉRENTES RÉGIONS

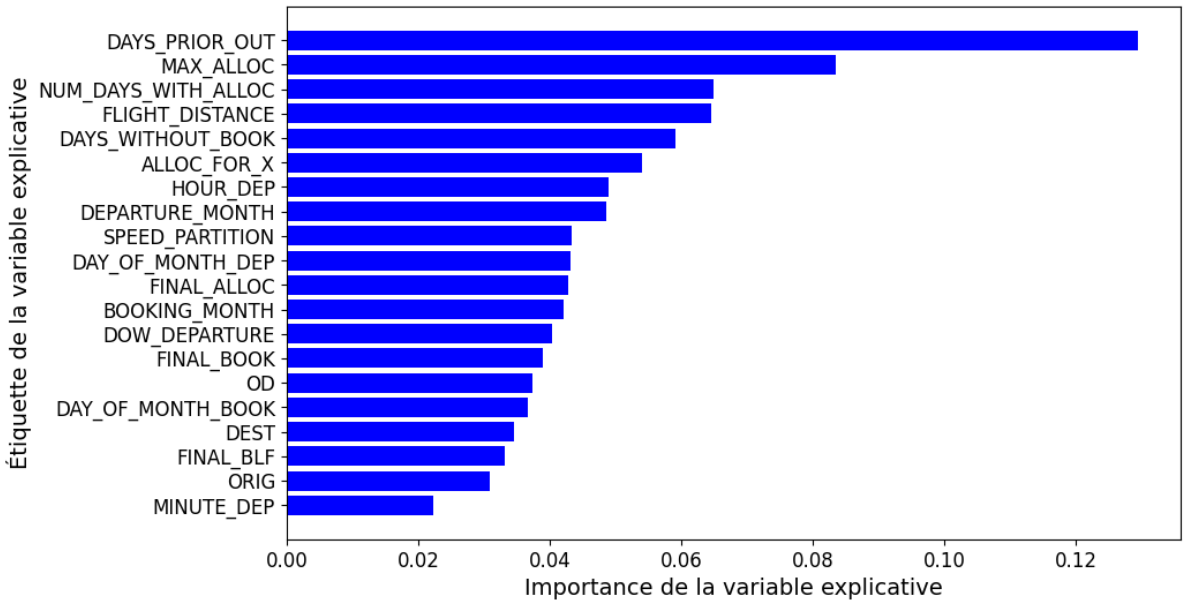


FIGURE C.1 Importance des variables explicatives concernant la région Domestique

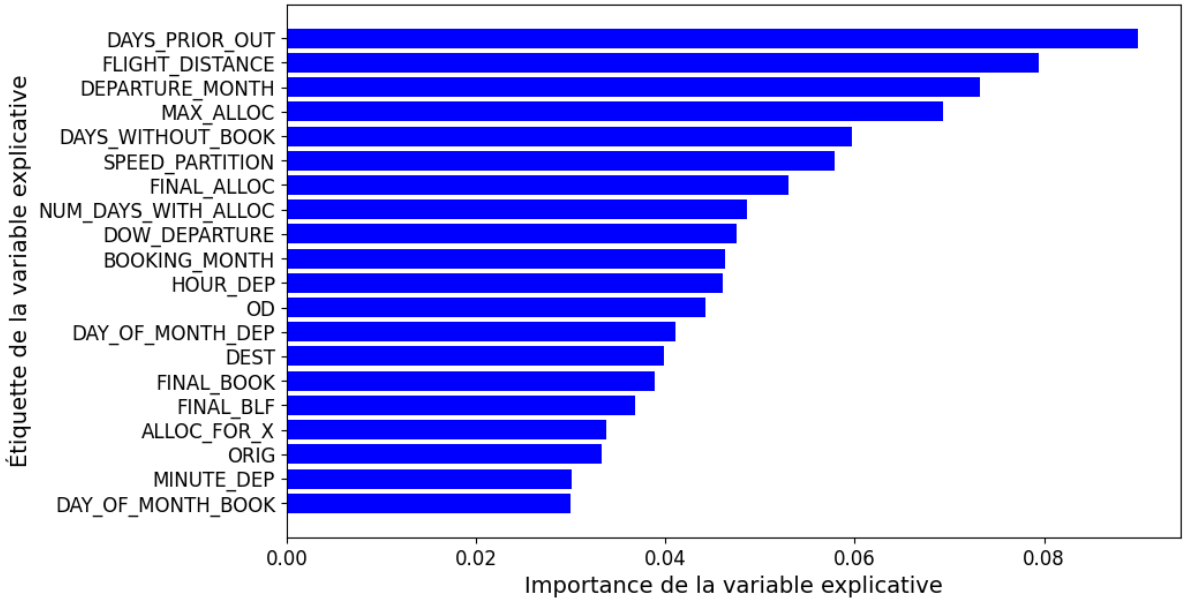


FIGURE C.2 Importance des variables explicatives concernant la région États-Unis

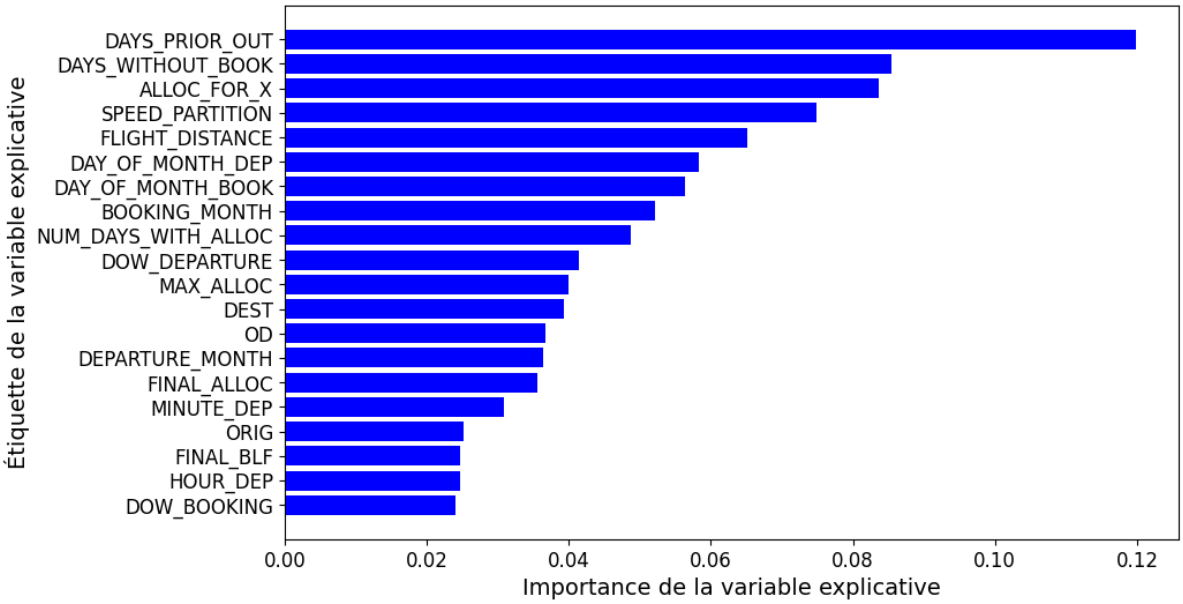


FIGURE C.3 Importance des variables explicatives concernant la région Pacifique

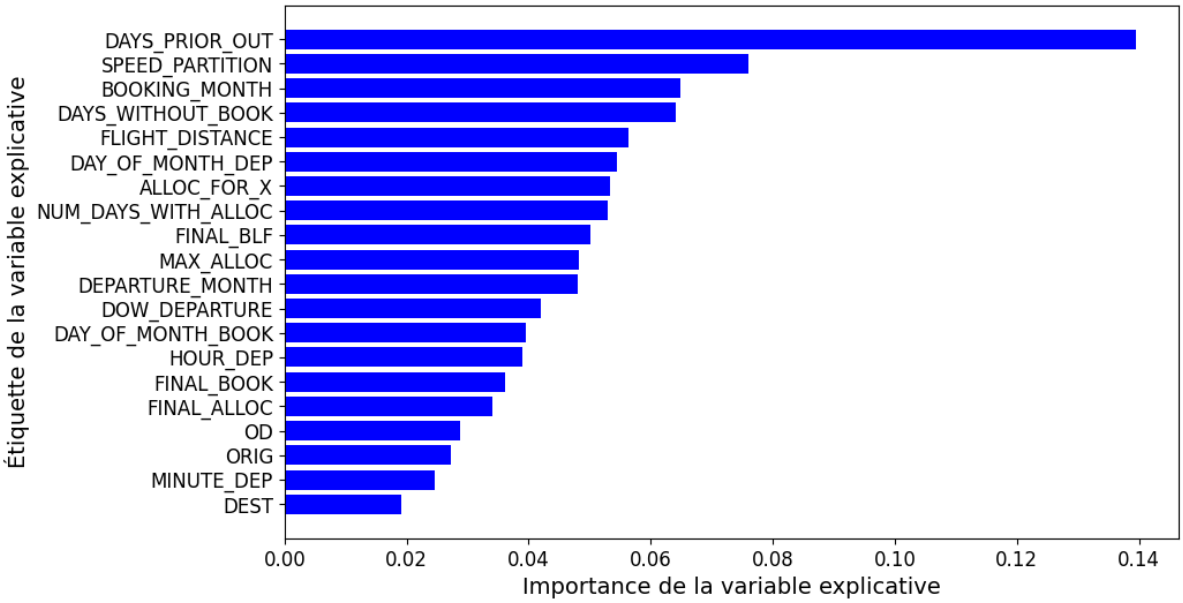


FIGURE C.4 Importance des variables explicatives concernant la région Atlantique

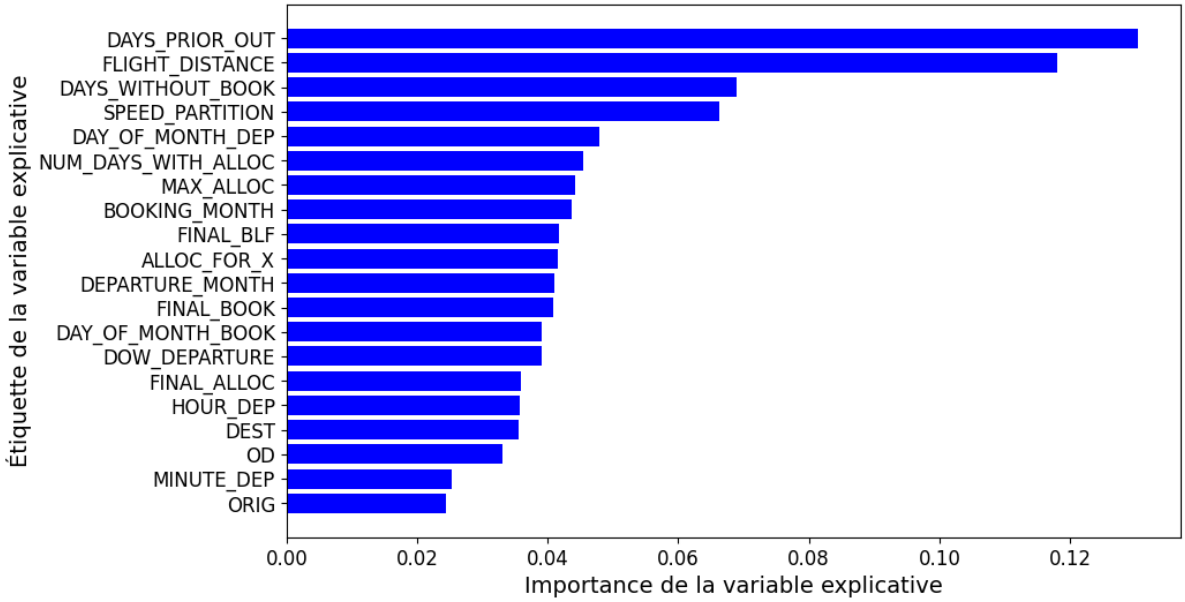


FIGURE C.5 Importance des variables explicatives concernant la région Sud