Titre: Title:	Modèles prédictifs de l'état final des devis de vente dans une grande entreprise de télécommunications
Auteur: Author:	Victor Devaux
Date:	2022
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Devaux, V. (2022). Modèles prédictifs de l'état final des devis de vente dans une grande entreprise de télécommunications [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/10268/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/10268/
Directeurs de recherche: Advisors:	Jean-Marc Frayret, & Luc Adjengue
Programme: Program:	Maîtrise recherche en génie industriel

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Modèles prédictifs de l'état final des devis de vente dans une grande entreprise de télécommunications

VICTOR DEVAUX

Département de mathématiques et de génie industriel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*

Génie industriel

Avril 2022

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Ce mémoire est intitulé :

Modèles prédictifs de l'état final des devis de vente dans une grande entreprise de télécommunications

présenté par **Victor DEVAUX**

en vue de l'obtention du diplôme de *Maîtrises ès science appliquées*

a été dûment accepté par le jury d'examen constitué de :

Camélia DADOUCHI, présidente
Jean-Marc FRAYRET, membre et directeur de recherche
Luc ADJENGUE, membre et codirecteur de recherche
Bruno AGARD, membre

REMERCIEMENTS

Je tiens à remercier vivement mes directeurs Monsieur le Professeur Jean-Marc Frayret et Monsieur le Professeur Luc Adjengue, Professeur à l'Université de Polytechnique Montréal, qui m'ont encadré tout au long de ce projet de recherche. Je leur suis reconnaissant de m'avoir fait bénéficier de leur grande compétence, de leur rigueur intellectuelle, et de leurs conseils avisés. Leur écoute et leur disponibilité ont été déterminantes dans l'aboutissement de ce projet. Soyez assurés de ma profonde gratitude. Je tiens également à remercier mes différents interlocuteurs au sein de l'entreprise partenaire et plus particulièrement François qui, par le temps qu'il a pu consacrer à nos entretiens, m'a permis de comprendre les enjeux et les particularités des processus de son entreprise. A titre plus personnel, je remercie chaleureusement ma sœur Oriane, pour ses encouragements et la grande patience dont elle a fait preuve à la relecture de mon manuscrit. Je tiens à remercier également mes parents et mon amie Émilie pour leur soutien moral ininterrompu et leurs encouragements précieux. Enfin, je voudrais exprimer ma reconnaissance envers mes amis qui m'ont apporté leur soutien tout au long de ma démarche.

RÉSUMÉ

La prévision de vente est un enjeu fondamental dans le domaine des hautes technologies. Ce dernier fait face à de nombreuses difficultés à cause du court cycle de vie de ses produits et de la forte concurrence. Ce mémoire présente une étude sur la possibilité d'utiliser le cycle de vie des devis pour améliorer les prévisions des ventes à court et moyen terme dans une entreprise de télécommunication. La gestion des devis est devenue un élément stratégique pour l'entreprise, car elle lui permet de limiter les surstocks liés aux déséquilibres entre les délais de production et de livraison. Les devis sont aujourd'hui gérés par un processus manuel. Ils sont évalués en les intégrant dans un processus de prévision long terme et ainsi prédire la demande réelle au sein d'un trimestre de vente. Cette étape est aujourd'hui une limitation dans le processus de prévisions de l'entreprise. La problématique soulevée a été la prédiction de l'état final des devis de l'entreprise. Pour cette étude, nous avons exploité les données de leur évolution sur 18 mois. Les analyses préliminaires ont déterminé qu'il existait différents types de clients et de devis. La création de modèles prévisionnels spécifiques à chaque client apparaissait donc envisageable. Nous avons donc développé deux modèles sur R : l'un basé sur des arbres de décisions et l'autre sur les forêts aléatoires. Les modèles ont été validés par un processus de validation croisée. L'exploitation de leurs résultats sur cinq clients a montré que la fiabilité des prédictions de vente dépendait du type de client. En fonction de celui-ci, l'erreur sur la justesse varie entre 10 % à 30 % et les scores F1 pour la classe « ANNULÉ » atteignent des valeurs entre 60 % et 80 %. La prédiction de l'évolution de l'état final des devis est donc possible pour une partie des clients. Cette démarche donne d'ailleurs des résultats prometteurs dans la prédiction de la vente, mais reste sujette à analyse dans celle d'éventuel retard dans la signature de la commande de vente. Nous démontrons ainsi que les données disponibles sur les devis devraient permettre à l'entreprise d'améliorer ses prévisions de ventes et ainsi réduire les surstocks. Toutefois, des analyses approfondies sont nécessaires pour comprendre en détail le processus de gestion de devis et proposer des mécanismes permettant de réduire la marge d'erreur pour les clients dont les prévisions sont les moins précises.

ABSTRACT

In the high-tech sector, sales forecasting is an essential element. This hyper-competitive sector experiences rapid product renewal. The management of quotes is strategic for the company because it allows limiting, upstream of the orders, the overstocks linked to the production and delivery delays. This thesis presents a study on the possibility of improving short- and medium-term sales forecasts in a telecommunication company based on the quote life cycle. Currently, in this company, quotes are managed by a manual process. To estimate the demand for a sales quarter, they are analyzed by the salespeople and then by a regulator, also considering a long-term forecast. This step has become a limitation in the forecasting process. The problematic raised was the anticipation of the final state of the quotes. For this study, we exploited data on the evolution of the company's quotes over 18 months. Preliminary analyses determined that there were several types of customers and quotes. The creation of specific forecasting models for each customer appeared feasible. We therefore developed two models in R. The first one is based on decision trees and the second on random forests. The models were validated by a cross-validation process. The exploitation of their results on the five main buyers showed that the reliability of the sales predictions depended on the type of customer. Depending on the type of customer, the predictions varied from 10% to 30% error on accuracy and F1 scores for the "CANCELLED" class between 60% and 80%. The prediction of the final evolution of the quotes is thus possible for a part of the customers. It gives promising results in the prediction of sales but remains to be deepened for that of possible delays in the signature of orders. We demonstrate that the data available on the quotes should allow the company to improve its sales forecasts and thus reduce the overstocks. Further analysis is needed to refine the understanding of the quote management process and to propose models to reduce the margin of error for customers with the least accurate forecasts.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABREVIATIONS

ADD	Arbre de Décision Détaillé
ADS	Arbre de Décision Simple
CRD	Date demandée par le client - Client Requested Date
DDCC	Données DC consolidées
FAD	Forêts Aléatoires Détaillée
FAS	Forêts Aléatoires Simple

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

La prévision de la demande est un défi stratégique dans la gestion d'une entreprise et notamment pour l'optimisation du fonctionnement de son réseau logistique. L'environnement économique incertain actuel [26] oblige les entreprises à innover toujours plus pour répondre au plus vite et sans erreur aux demandes clients [10]. Les produits technologiques sont particulièrement concernés par cette problématique, car manufacturés en flux poussé. Il est donc essentiel de prévoir au plus juste les futures ventes pour éviter tout retard d'approvisionnement ou de surproduction.

1.1 Contexte général de l'entreprise partenaire

Le partenaire de cette étude est une entreprise de haute technologie qui gère l'innovation, la conception et la vente de produits de haute technologie. Il s'agit d'un acteur majeur de son secteur. Elle possède une large gamme de produits qui sont vendus dans le monde entier à de nombreux clients. L'entreprise dispose d'un réseau mondial de vente dans lequel les clients sont divisés en région et sous-région.

L'entreprise partenaire fonctionne en « contract manufacturing ». Autrement dit, elle développe de nouvelles technologies dont elle possède les brevets, mais la production est sous-traitée. Pour une entreprise souhaitant livrer avec des délais courts dans le monde entier, la sous-traitance représente donc un vrai défi organisationnel. La logistique d'acheminement des produits finis aux clients reste gérée par l'entreprise. Plusieurs lieux de stockage existent pour permettre une livraison dans des délais optimaux, quel que soit le produit demandé. La prévision de la demande est donc un enjeu clé pour l'entreprise. Une bonne prévision lui permettra d'éviter les surstocks ou les ruptures de stock qui pourraient occasionner la perte de contrats dans un milieu toujours plus compétitif.

1.2 Vente, devis et prévision de ventes chez le partenaire

Cette sous-partie présente le processus de prévision chez l'entreprise partenaire et ses spécificités. Les prévisions de vente se basent sur les ventes réalisées et potentielles. Les clients, les vendeurs et les responsables logistiques contribuent au calcul des prévisions à travers différentes étapes.

1.2.1 Prévision chez le partenaire

Chez l'entreprise partenaire, la prévision de la demande globale fait l'objet d'un suivi en dollars d'équipements vendus ou prévisionnels par quart. Un quart, ou trimestre correspond à une période de 3 mois consécutive. Par exemple, le premier quart de l'année 2019 commence le 1er novembre 2018 et finit le 31 janvier 2019. Le suivi à l'intérieur d'un quart est réalisé par mois. Pour un quart donné, le suivi est réalisé de 3 mois avant le début du quart, jusqu'à la fin de celui-ci, soit sur une durée totale de 6 mois. Afin de faire une prévision globale des ventes en dollars, un suivi des prévisions des ventes et des commandes est réalisé par produit et par région. Chaque région est gérée individuellement. Au sein d'une région, les prévisions sont calculées par produit ou regroupement de produits. Cela permet au département logistique de suivre les commandes réelles et prévoir la demande globale au sein d'un même quart. Ce suivi permet également d'évaluer la qualité des prévisions calculées en début de quart.

La section suivante détaille ce processus de prévision et introduit plus spécifiquement la problématique générale de recherche.

1.2.2 Processus de prévision

Le fonctionnement du processus de calcul des prévisions des ventes par un quart au sein de l'entreprise est spécifique au partenaire. Ce calcul est basé sur un mélange de prévisions à long terme ainsi que sur l'utilisation de données relatives aux devis de vente déjà créés.

1.2.2.1 Ventes et devis

Chez le partenaire du projet, la vente des produits peut se dérouler de deux manières distinctes. Les clients peuvent faire des achats avec ou sans devis. La vente par devis représente environ 63 % des ventes. La vente sans devis ne sera pas étudiée dans ce projet.

Les devis suivent un processus et des cycles de vie précis. Ils sont créés par les vendeurs sur demande du client, et contiennent plusieurs informations, comme les produits demandés, la date de livraison souhaitée, le nom du client, et surtout un indicateur de confiance dans la réalisation du contrat. Le devis évolue alors à travers différentes étapes de négociation avec le client qui peut réviser certains aspects de sa demande. L'indicateur de confiance dans la réalisation du contrat peut être révisé si le vendeur le juge nécessaire. Si cet indicateur est assez élevé, le devis est pris

en compte dans le processus de prévision. Si la vente se concrétise, le devis se transformera en une ou plusieurs commandes. Dans le cas contraire, il sera annulé ou reporté au quart suivant. Durant ces étapes de négociation avec le client, plusieurs types d'événements peuvent affecter le devis :

- Le panier de produits peut évoluer sur demande du client ;
- La confiance du vendeur dans le devis peut évoluer à la hausse ou à la baisse ;
- La date souhaitée de livraison par le client peut être modifiée.

1.2.2.2 Prévision et devis

La Figure 1.1 présente le principe général du processus de calcul des prévisions des ventes pour la fin d'un trimestre. Tout d'abord, deux niveaux de prévision sont pris en compte par l'entreprise : une prévision à long terme, et une prévision à court terme avec l'utilisation des devis.

Les prévisions générales correspondent aux prévisions longs-termes du partenaire. Elles ne seront pas étudiées dans ce mémoire. Les prévisions courts termes sont réalisé à l'aide des devis et des prévisions longs-termes.

Les devis sont utilisés manuellement pour affiner les prévisions des ventes au sein d'un quart. Comme discuté plus haut, un devis est créé à la suite d'une négociation entre un client et un vendeur. Cette négociation permet de définir le besoin à un instant du client. Le vendeur intègre à son devis à un indicateur de confiance dans la réalisation de la vente pour le quart souhaité par le client. Cet indicateur prend la forme d'une probabilité estimée qualitativement par le vendeur, et pouvant être révisée en tout temps. Lorsque le devis atteint une valeur seuil, celui-ci intègre le processus de prévision.

Pour obtenir la prévision pour un quart de vente, le département logistique de l'entreprise utilise les données courantes des devis et les prévisions générales calculées par un autre département, ce qui permet d'obtenir une prévision brute. Cette prévision brute est ensuite analysée par le département logistique dans un but de raffiner les prévisions. L'analyse est faite manuellement par un analyste qui utilise ses connaissances des clients et des vendeurs. Il peut par exemple diminuer l'indicateur de confiance et donc les prévisions de vente pour un client et un vendeur

qu'il sait trop optimiste. Cette analyse des devis utilise ici une connaissance implicite des comportements clients et vendeurs.

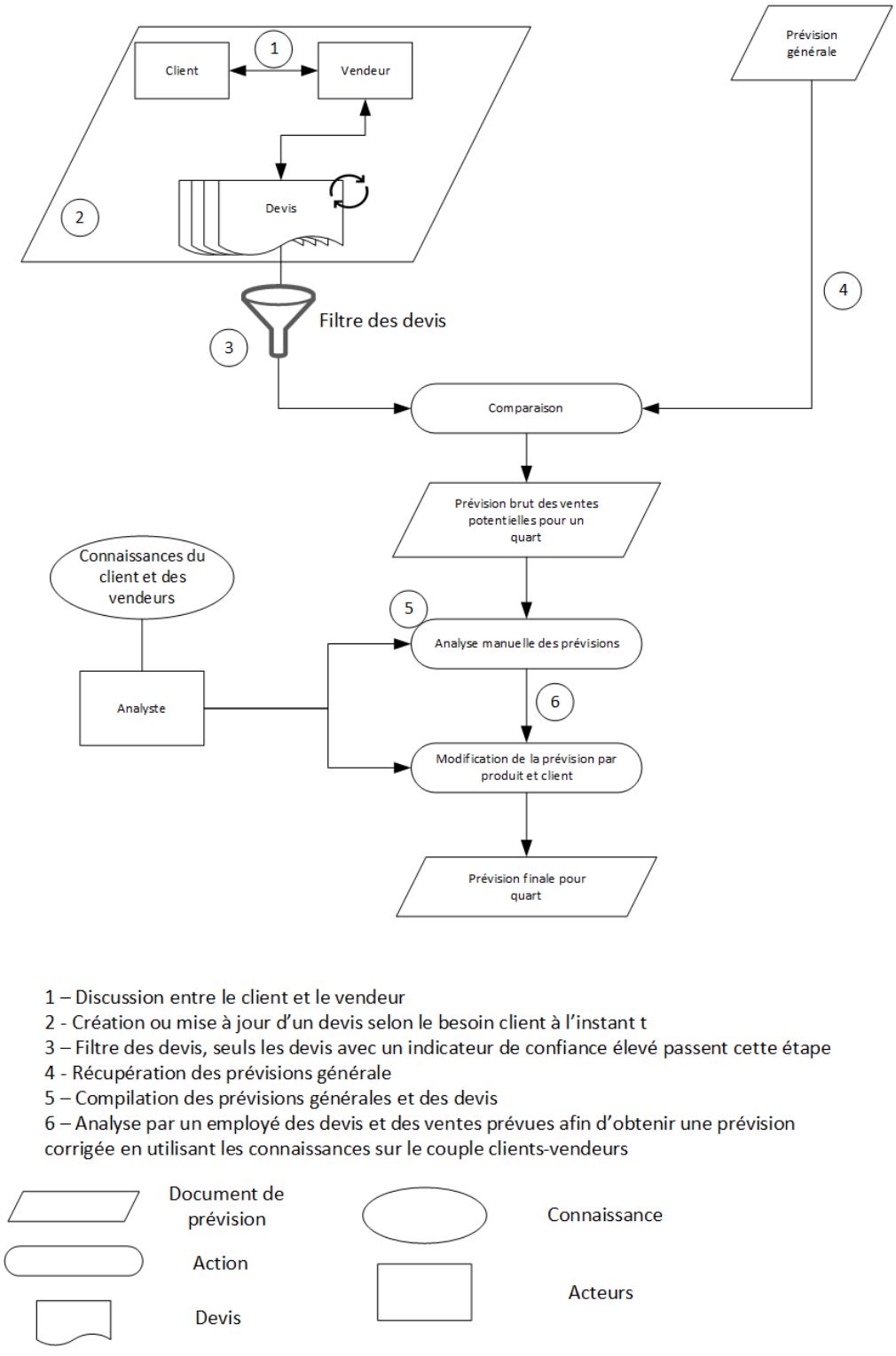

Figure 1.1 : Processus de prévision chez le partenaire

1.3 Problématique de recherche

La gestion des devis est enjeu pour l'émission des prévisions du partenaire. La gestion des devis est complexe et chronophage. Les devis créés sont filtrés par un processus d'évaluation subjective. Ils ne sont pris en compte par le département logistique qu'après que le vendeur n'ait qualifié leur indicateur de confiance comme étant suffisant. Toutefois, la fiabilité de cet indicateur reste limitée, car il n'est attribué que sur l'évaluation subjective du vendeur. C'est pourquoi une analyse manuelle est effectuée a posteriori par le département logistique si une commande paraît inhabituelle.

Les devis des clients évoluent dans le temps. Entre le début du projet et sa réalisation, un client peut changer d'avis pour différentes raisons comme l'évolution technique de certains composants faisant évoluer les spécifications du devis. Les produits étant très spécifiques et d'un haut niveau technologique, les achats de certains composants ne sont pas réguliers. Selon les plans d'investissement de leurs clients, la demande pourrait être très élevée ou non. Cette demande intermittente complique les prévisions de vente. De plus, un devis, contrairement à une commande, n'est pas engageant. Le client peut annuler sa demande à tout moment. Savoir déchiffrer les caractéristiques et spécificités des devis et des clients peut s'avérer utile pour anticiper l'évolution des devis et ainsi améliorer les prévisions de la demande. On peut se fier, par exemple, sur des comportements systématiques des clients, leur niveau de fiabilité, ou des similarités entre les clients.

Nous chercherons à savoir si l'analyse des caractéristiques des devis notamment de leurs différentes évolutions, permet de prédire la réalisation de ceux-ci et ainsi remplacer ou a minima compléter les analyses manuelles actuellement en place (Figure 1.2).

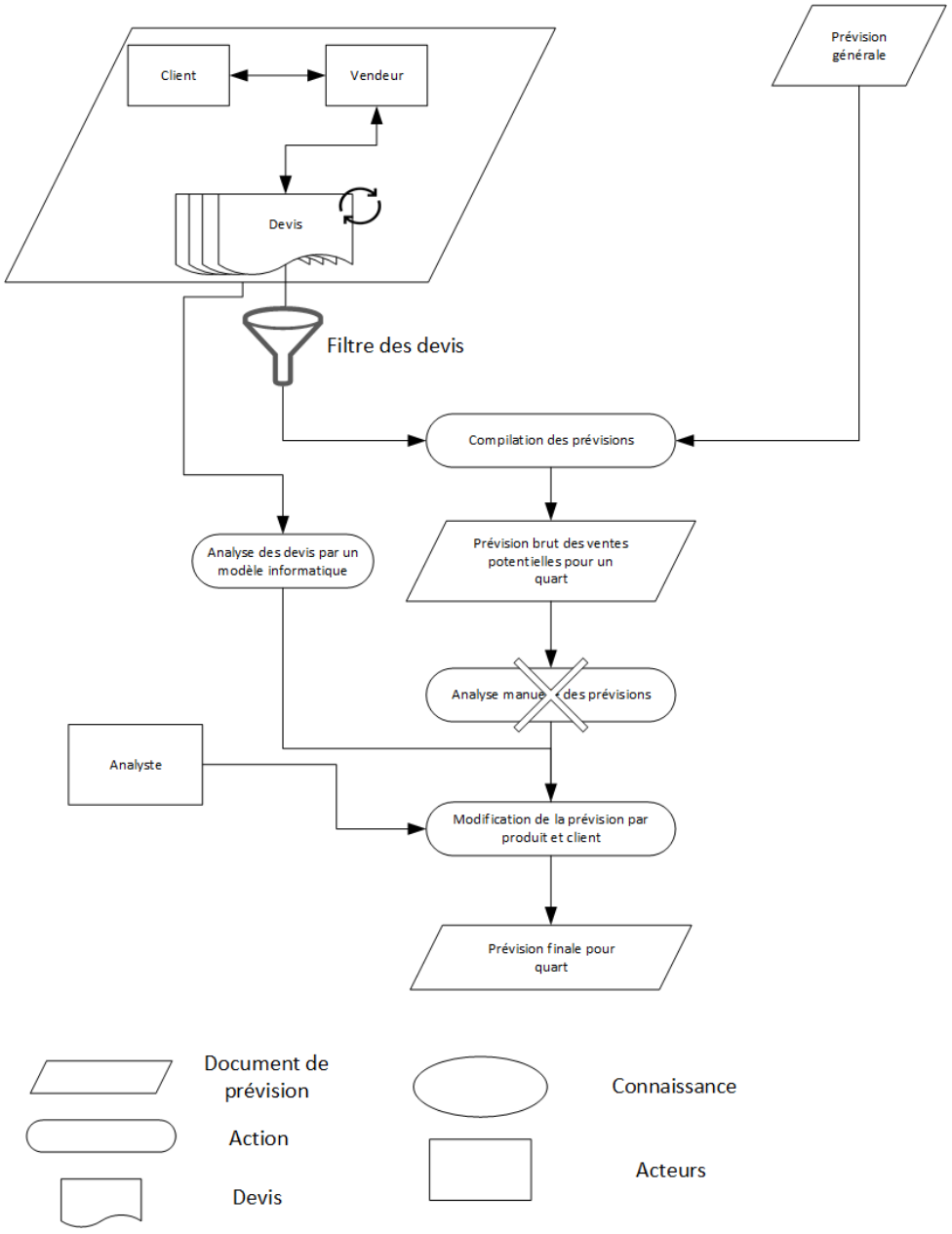

Figure 1.2 : Évolution potentielle du processus

CHAPITRE 2 ANALYSE DE LA LITTÉRATURE

2.1 Introduction

L'objectif de cette analyse de la littérature est de faire un état de l'art dans le domaine de la prévision des ventes et de la demande. Nous étudierons plus particulièrement les outils de prévision des ventes et de la demande en nous basant sur l'utilisation d'informations disponibles sur les devis, ainsi que les méthodes de classification, et finalement les méthodes d'analyse de séries temporelles.

La structure de cette revue est la suivante. Dans un premier temps, nous exposerons la méthodologie de recherche des articles. Ensuite nous présenterons une analyse de la littérature existante en insistant sur les méthodes actuelles et les facteurs complexes spécifiques au domaine de la haute technologie. Enfin, nous proposerons des opportunités de recherche en comparant les articles à la problématique décrite précédemment.

2.2 Méthodologie de recherche d'articles

Dans le cadre de cette revue de littérature, nous avons utilisé une méthode de recherche bibliographique systématique. L'analyse de la littérature s'est organisée en deux temps :

1. **Définition de la problématique et des principaux mots clés** : cette étape a été réalisée à l'aide du moteur de recherche *Google Scholar* et des recherches concentrées sur des états de l'art publiés.

Les résultats de *Google Scholar* nous ont permis de préciser les concepts spécifiques à la prévision de la demande et les termes récurrents associés dans la littérature. Cette analyse préliminaire avait pour objectif d'identifier les termes pertinents du sujet étudié et de définir les mots clés à utiliser dans la recherche approfondie d'articles. Les mots clés ont été sélectionnés à la suite de ces recherches et de la spécificité du partenaire qui œuvre dans le domaine de la haute technologie. Cette analyse nous a permis d'identifier différentes familles de mots clés, présentées dans le Tableau 2.1. La colonne Industrie permet d'étudier le milieu de l'entreprise partenaire.

Tableau 2.1 : Familles de mots clés

Prévision	Demande	Méthodes	Industrie	Devis
Forecast* Prevision	Demand Sales diffusion	Times series Classification	Industries High-Technology Telecom* semiconductor	Quote Quotes

2. **Identification d'articles pertinents dans Compendex**

Ensuite, la recherche systématique par mots clés dans le sujet, titre ou résumé a été réalisée sur la base de données Compendex entre 1995 et 2022. Seuls les articles de conférences et les articles de journaux ont été retenus. Le critère de langue anglaise a aussi été sélectionné. Des critères d'exclusion ont également été mis en place (Tableau 2.2).

Tableau 2.2 : Critères d'exclusion des articles

Critères d'exclusion	Description
E1	Le mot « quote » utilisé dans le sens de citation
E2	Articles non disponibles ou payants
E3	Articles hors sujet
E4	Articles en doublons
E5	Domaine non pertinent NOT (corporate* OR financial* OR electric* OR energ* OR water OR commerce OR traffic OR marketing)

Les mots clés identifiés ont été combinés afin de définir les recherches spécifiques présentées dans le Tableau 2.3 :

Tableau 2.3 : Recherches utilisées dans Compendex

Recherche 1	(demand OR sales) AND (forecast* OR predict*) AND (quotes AND quote)
Recherche 2	(Demand OR Sales) AND (forecast* OR predict*) AND (semiconductor OR telecom*) and times series)
Recherche 3	(demand OR sales OR diffusion) AND (forecast* OR predict*) AND Industr* AND (semiconductor OR telecom*) AND (times series OR classification)
Recherche 4	(demand OR sales) AND (forecast* OR predict*) AND Industr* AND semiconductor

La Recherche 1 cherche permet d'identifier la littérature qui apporte des solutions à la prévision par devis ou à leur utilisation dans des processus de prévision des ventes. Les Recherches 2 à 4

sont plus larges. Elles ont pour objectif d'identifier des articles concernant la prévision à travers différentes méthodes, dans des domaines proches du cas d'étude.

Les recherches réalisées ont permis d'obtenir 686 articles. Sur ces 686 articles, 325 ont été retenus après exclusion de ceux dont les domaines d'études étaient non pertinents (critère d'exclusion E5). Sur ces 325 articles retenus, 29 ont été étudiés après application des critères d'exclusion présentés plus haut à la suite de la lecture des résumés (Tableau 2.4).

Tableau 2.4 : Critères d'exclusion par recherche

Recherche	Nombre d'articles avant exclusion	Critères d'exclusions appliqués	Nombre d'articles retenus
1	71	E1, E2, E3	5
2	115	E2, E3, E5	5
3	118	E2, E3, E4, E5	9
4	382	E2, E3, E4, E5	9

Lorsque ces 29 articles ont été identifiés, nous avons poursuivi la recherche manuellement avec la méthode du « backward and forward snowball sampling » définie par le Docteur Claes Wohlin [33]. Cette dernière s'appuie sur la lecture des études citées dans les références ou les bibliographies des articles de l'échantillon initial. Cette méthode a permis d'identifier 4 articles pertinents supplémentaires. Finalement, 32 articles ont été retenus. Le Tableau 2.5 présente une synthèse de ces articles. La méthodologie de recherche décrite est résumée dans la Figure 2.1.

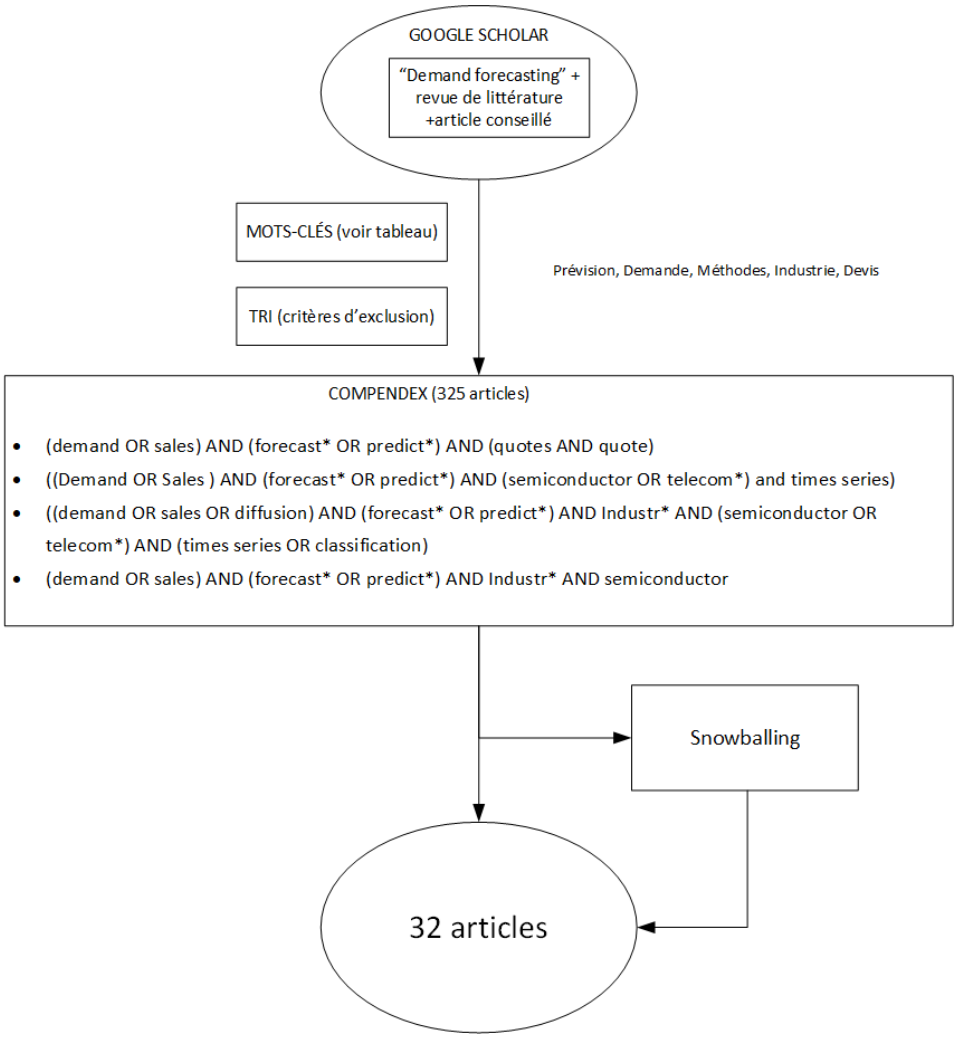

Figure 2.1 : Méthodologie de la revue de littérature

2.3 Prévision de la demande

2.3.1 Importance et utilité de la prévision

La prévision de la demande est un processus clé dans toutes les industries [32]. Basée sur une analyse des données historiques [10], elle s'inscrit plus particulièrement dans les problématiques d'aide à la décision [12] à différentes échelles de temporalités [20]. Elle permet par exemple d'optimiser la réponse aux commandes des clients [1]. De nombreuses décisions stratégiques sont basées sur la prévision de la demande [6], car elle impacte directement la planification des affaires et la stratégie commerciale des entreprises [24]. En effet, elle fournit des données

essentielles pour soutenir les décisions stratégiques de planification et les dépenses d'investissement associées qui nécessitent de longs délais d'exécution [28,31].

L'utilisation de bonnes prévisions permet aussi aux entreprises d'améliorer leurs processus et donc leur rentabilité [11]. Grâce à ces prévisions, elles peuvent anticiper les comportements clients [22] et devenir plus agiles, réactives et flexibles [13]. Par exemple, les entreprises peuvent temporiser l'effet coup de fouet [32] qui consiste en une amplification importante des variations de la demande qui croît avec la longueur de la chaîne d'approvisionnement [21,26]. À l'inverse, si une entreprise n'a pas recours à de bonnes prévisions, cela peut causer une perte de qualité de service et donc une perte de contrat [1,4,18,19,26].

2.3.2 Spécificité du milieu technologique

La prévision dans le domaine de la haute technologie est généralement plus difficile que dans les industries traditionnelles [9]. Une bonne prévision est un facteur de différenciation clé dans le secteur. Par exemple dans le milieu des semi-conducteurs qui ont une durée de vie limitée, anticiper la demande au plus juste permet d'éviter la surproduction qui entraînerait des pertes importantes si le produit devenait obsolète [8]. Mais cela permet également d'éviter la sous-production de produits à haut coût d'inventaire, qui conduirait à une rupture de stock et donc à un manque à gagner [8]. Par ailleurs, si le produit rencontre un succès rapide, il faut être en mesure de répondre rapidement à la demande et donc d'accélérer sa fabrication, qui, pour les semi-conducteurs, nécessite parfois la construction coûteuse d'usines spécifiques [26].

Le secteur fait donc face à de nombreux défis [8,23]. Tout d'abord, le marché des produits de hautes technologies est très compétitif [26]. [23] soulignent que la concurrence croissante avec l'Asie présente un défi dans ce secteur. Un autre défi de ce secteur est la forte volatilité de la demande [9,16,30,31] et les croissances rapides de la demande qui sont parfois observées [31]. Enfin, le marché fait face à un manque d'information des acteurs en aval de la chaîne d'approvisionnement et à une intermittence de la demande [8]. Une demande intermittente est définie dans l'article [8] comme une demande qui apparaît de manière aléatoire avec un pourcentage élevé de valeurs nulles entre les occurrences de demande non nulle. Pour rester compétitifs, les fabricants doivent être en mesure d'offrir à leurs clients un niveau élevé de flexibilité dans leur commande afin de les aider à s'adapter à leurs consommateurs [23].

La fabrication des produits de haute technologie et la gestion de leur durée de vie sont par ailleurs très complexes [10]. La difficulté des prévisions vient de leur coût important [8], du nombre [8] et du degré de personnalisations [23] et surtout de leur courte durée de vie [20]. En effet, cette dernière est de plus en plus courte [20] à cause de la rapidité des mises à jour ou de la création de substituts [10]. Leur durée de vie est comprise entre 1,5 et 2 ans [10]. Il devient alors difficile à cette échelle de temps, à cause du peu de données historiques, de faire des prévisions fiables. [15].

2.4 Méthodes de prévisions de la demande

Pour faire face aux difficultés de prévision, de nombreuses méthodes ont été développées. Les modèles de prévision utilisent des modèles statistiques, mathématiques ou empiriques pour estimer la demande future [6].

2.4.1 Différentes méthodes et facteurs

2.4.1.1 Des méthodes quantitatives et qualitatives

De nombreuses méthodes de prévisions existent dans la littérature. Les professionnels du secteur utilisent des méthodes quantitatives et qualitatives [9,17]. Les méthodes de prévision qualitatives reposent sur le jugement et la théorie d'experts [9,17]. Ces prévisions sont plutôt pertinentes pour des entreprises avec peu d'historiques. Les méthodes de prévision quantitatives comprennent par exemple l'analyse de régression, l'analyse des séries chronologiques, le lissage exponentiel ou les réseaux neuronaux [9].

Les méthodes quantitatives consistent à développer des modèles statistiques, mathématiques ou empiriques pour estimer la demande future [6,28]. Ces modèles reposent principalement sur la disponibilité de données et d'informations passées lors de leur construction [6,10,17].

La quantité de données utilisables augmentant [29], elles ne peuvent plus être uniquement traitées à l'aide des techniques et d'applications informatiques classiques [15]. Face à cette problématique, l'exploration de données permet d'identifier les tendances ou les modèles cachés dans un ensemble important de données [18] par des moyens automatiques ou semi-automatiques [29].

De plus [18] indiquent que des explications sont cruciales pour l'acceptation de ces outils de prévision, et ainsi qu'un bon modèle doit répondre à cinq critères : bonne compréhension, fidélité, précision, évolutivité et généralisation.

Les méthodes qualitatives et quantitatives se complètent toutefois. Les avis d'experts permettent de compléter les données statistiques en tenant compte des évolutions conjoncturelles d'une industrie qui pourraient ne pas être visibles dans le traitement des données. Il faut toutefois disposer d'indicateurs pertinents issus de l'analyse quantitative pour ne pas fausser l'analyse qualitative [5].

2.4.1.2 Choix des données et facteurs

Pour développer ces modèles, différents facteurs sont pris en compte selon les bases de données disponibles et le ou les modèles choisis. Un des premiers facteurs est le type de prévision que l'on souhaite réaliser. La littérature développe deux types de modèles : de la classification [16] ou de la prévision par régression [3]. Le développement de modèle dépend également de la temporalité : prévision court terme ou long terme [2,24]. Les modèles peuvent nécessiter également un volume de données disponibles plus ou moins important pour être efficaces [9]. [8] indique que la sélection des données d'entrée est importante pour le développement d'un bon modèle.

Le choix des facteurs a une influence importante sur la prévision finale. Certains articles étudiés proposent des prévisions avec peu de facteurs externes alors que d'autres intègrent un grand nombre de ces facteurs. On peut séparer les facteurs étudiés en plusieurs catégories :

- Les facteurs liés aux caractéristiques des produits comme l'historique des ventes, le prix des produits [32], ou le cycle de vie des produits. En effet, le cycle de vie des produits dans le secteur des hautes technologies suit régulièrement un modèle classique de la théorie de l'innovation [17,32]. Autrement dit, leur évolution passe à travers plusieurs étapes : introduction, croissance, maturité et déclin [20].
- Les facteurs liés aux comportements des clients [14] : la saisonnalité des achats [32], leurs biais (la sous-estimation ou surestimations constante des commandes [27]), et la confiance dans la relation client-vendeur [21,27].

- Les facteurs liés au marché comme les facteurs environnementaux [7], les conditions économiques, les parts de marché attendues ou marché total disponible [24].

2.4.2 Modèles basés sur l'analyse de séries chronologiques

Cette sous-partie est consacrée à la présentation des différents modèles présents dans la littérature et leur capacité de prise en compte des facteurs cités dans la partie précédente. La classification et la prédiction sont deux formes d'analyse des données. Elles peuvent être utilisées pour extraire des modèles décrivant les classes de données importantes ou pour prédire les tendances futures des données [29].

2.4.2.1 Prédiction des séries chronologiques

2.4.2.1.1 Modèles classiques et modèles autorégressifs

Dans cette partie, une première sous-catégorie est constituée des modèles de lissage (Smoothing) : lissage exponentiel (simple ou double) [3,9], qui ne fait intervenir aucune notion de variables aléatoires. Une deuxième sous-catégorie est constituée de modèles stochastiques : modèles autorégressifs (AR), moyenne mobile (MA) et plus généralement ARMA (Moyenne mobile autorégressive) [3,6,25] et ARIMA (Moyenne mobile autorégressive intégrée) [6,13,24]. Ces deux sous-catégories de modèles sont très utilisées, car faciles à mettre en place. Ils utilisent comme données d'entrée celles issues des historiques de ventes. Le lissage exponentiel n'est pas efficace dès que les données se compliquent et ils ne peuvent pas prendre en compte des facteurs d'entrée importants comme les devis ou la saisonnalité [3]. Il ne s'agit donc pas des modèles les plus performants dans notre domaine d'étude. Toutefois les modèles ARMA, ARIMA ou les modèles vectoriels ARMA (ARMAV) [3] ou VAR [13] permettent d'améliorer des prévisions. Ces modèles ont un fonctionnement commun. Ils prennent ainsi en compte les données historiques et les décomposent en un processus autorégressif (AR), dans lequel il existe une mémoire des événements passés. ARIMA dispose d'un processus intégré (I) qui permet de stabiliser ou de rendre les données stationnaires, ce qui facilite la prévision. De plus, ARIMA présente une moyenne mobile (MA) des erreurs de prévision, de sorte que plus les données historiques sont longues, plus la prévision sera précise. Enfin, les modèles ARIMA, ARMAV et VAR apprennent avec le temps [3,13,24]. La différence entre les modèles VAR et ARIMA est que VAR suppose que les relations peuvent être approximées en utilisant uniquement les

composantes autorégressives. Ces modèles peuvent aussi être complexifiés en prenant en compte la saisonnalité [3,6], ou l'historique des devis sur 3 mois [3]. L'article [3] explique cependant que la seule prise en compte d'une tendance n'est pas suffisante. Il préconise l'utilisation du modèle ARMAV avec la prise en compte de tendances et de facteurs externes pour obtenir des résultats optimaux dans le domaine de la haute technologie.

2.4.2.1.2 Modèles complexes

Il n'est toutefois pas toujours possible d'utiliser des séries chronologiques comme dans certains modèles autorégressifs. En effet, leur efficacité peut être limitée [20] à cause de relations non linéaires entre les données d'entrée. L'apprentissage machine peut ainsi être une solution pertinente à cette problématique. Il permet de réaliser des prédictions efficaces lorsque les modèles classiques ne sont plus applicables. Des modèles plus complexes sont proposés dans la littérature tels que les réseaux de neurones [5, 9, 20, 24] ou les modèles de prévision gris roulant (rolling grey forecasting model) [9,28].

Les réseaux de neurones sont des modèles complexes qui montrent une certaine efficacité dans le secteur des hautes technologies [6,8]. Ils peuvent prendre en compte plusieurs paramètres simultanément, s'adapter à la demande intermittente et aider la gestion des stocks par exemple [8]. Cependant [24] explique qu'ils présentent aussi certains défis, comme la non-prise en compte de la saisonnalité et la difficulté d'interprétation.

Les modèles gris sont des modèles qui possèdent de nombreux avantages. Ils ont souvent été étudiés dans l'industrie de la haute technologie. Un modèle de prévision gris roulant se reconstruit et se met à jour au fur et à mesure que de nouvelles observations sont disponibles. Les modèles gris glissants nécessitent peu de données [9,28] en comparaison à d'autres modèles complexes tels que les réseaux de neurones. Ceci est particulièrement intéressant pour le secteur où les produits changent rapidement et où la quantité de données historiques est donc faible [9]. Il peut de plus prendre en compte de nombreux facteurs ayant chacun un coefficient d'importance. Un inconvénient de ce modèle réside dans la difficulté du choix des coefficients et des facteurs à prendre en compte et de l'impossibilité de tenir compte des corrélations entre les indicateurs importants [9].

2.4.2.2 Classification, diffusion et modèle hybride

Un autre type de prévision est la prédiction par classification. Le modèle de classification est l'un des modèles les plus utilisés pour analyser les tendances et planifier l'avenir, par exemple dans les affaires[18].

2.4.2.2.1 Utilité

La littérature présente différents types d'éléments classifiables : les clients [1,27], les opportunités [4], les produits [10] ou encore les temps de production [16]. La classification de données permet d'améliorer les modèles de prévision par séries chronologiques mentionnées ci-dessus. La classification des clients permet par exemple en cas de pénurie de réserver des ressources aux clients à forte rentabilité [1] ou aux plus fiables [27]. La classification des clients peut aussi permettre d'identifier quels clients sont le plus susceptibles de partir [11,12,15,19]. Le regroupement de produits [13] permet par exemple de classer les produits ayant un même cycle de vie et ainsi adapter le modèle à un type de produit.

2.4.2.2.2 Modèles par classification

Différents modèles ont été développés pour réaliser des regroupements et des classifications. Les différents modèles étudiés dans la littérature sont la régression logistique [11], les arbres de décisions [16,18,19], le modèle naïf de Bayes [18], les forêts aléatoires [4,12,18,19], la méthode des k plus proches voisins [12] et les réseaux de neurones [15,16]. Ces articles montrent que la meilleure classification est souvent obtenue avec le modèle des forêts aléatoires.

2.4.2.2.3 Modèles de diffusion

Les modèles de diffusion sont aussi très utiles pour la classification. Ils sont régulièrement utilisés dans le secteur des hautes technologies, car l'innovation y est forte et donc l'obsolescence rapide [32]. L'analyse des cycles de vie permet de regrouper les produits qui présentent des similitudes et ainsi de prévoir la courbe des ventes de nouveaux produits dont on ne dispose pas de données historiques [32]. Un exemple de modèle souvent présenté dans la littérature est le modèle de Bass [17,20,32]. Le modèle de diffusion de Bass décrit le processus d'adoption de nouveaux produits comme une interaction entre les utilisateurs et les utilisateurs potentiels [17]. En particulier [20] et [32] développent des modèles multigénérationnels, qui utilisent les facteurs

suivants : la saisonnalité, le prix, l'évolution du marché, la répétition d'achat et l'effet de substitution technologique. [20] utilise aussi le cycle de vie produit comme facteur. Ces deux modèles permettent d'obtenir de bons résultats.

2.4.2.2.4 Modèles hybrides

Les modèles précédents peuvent être plus efficaces s'ils sont utilisés en complément d'un autre modèle. [24] affirment ainsi qu'une prévision devrait prendre en compte plusieurs modèles. [25] soutient également cela en utilisant un modèle vectoriel autorégressif conditionnel généralisé hétéroscédasticité (GARCH). Ce modèle exploite une combinaison de plusieurs modèles avec des poids optimisés pour une meilleure prévision.

2.5 Opportunité de recherche

Les prévisions des ventes dans le milieu de la haute technologie sont assez complexes comme nous avons pu le voir dans la partie précédente. C'est pourquoi les entreprises ont besoin de mieux comprendre le comportement des clients et du marché grâce à l'historique des ventes.

2.5.1 Problématique industrielle et données des devis

La problématique du partenaire de ce projet concerne une utilisation de données inusitées dans la littérature. On ne retrouve que deux articles qui prennent en compte les devis dans leur processus de prévision [3,4]. Les différentes bases de données utilisées pour réaliser des prévisions sont en majorité composées d'historiques de ventes ou d'informations clients [1,21].

Les données étudiées ici comprennent des « photos » des caractéristiques des devis prises à un intervalle de temps régulier. Ce type de données n'a pas été observé dans notre analyse de la littérature. De plus, ces données sont uniques, car elles sont spécifiques et n'ont jamais été utilisées pour cette fin par l'entreprise.

L'utilisation des devis, et particulièrement la prévision du cycle de vie des devis, ne semble jamais avoir été utilisée pour caractériser les comportements des clients, et ainsi anticiper la demande. Le Tableau 2.5, expose le fait que la littérature se concentre davantage sur l'analyse des séquences de vente par régression. Les quelques articles disponibles sur les modes de classification des commandes se concentrent sur des temporalités de prévision différentes du cas étudié dans ce mémoire. Ainsi, les seuls articles mentionnant la prise en compte de devis dans

leur modèle de prévision sont les articles [3] et [4]. Le premier choisit en facteur d'entrée les devis pour effectuer son modèle autorégressif. Toutefois, il n'aborde pas le concept d'acceptation ou non de la vente qui est étudiée dans [1]. [4], quant à lui, se concentre davantage sur l'analyse du comportement client. Par exemple, il prend en compte le temps de lecture de devis pour évaluer l'opportunité de vente. Il présente cependant le défaut d'avoir à demander beaucoup d'informations au client [4].

Malgré ce manque d'intérêt pour les données de cycle de vie des devis, il est possible de comparer la prévision de la vente de devis avec les classifications de désabonnement des clients [11, 12, 18, 19]. En effet, les données, les objectifs et les produits utilisés présentent des similitudes entre les deux situations étudiées. Dans les deux cas, les données sont déséquilibrées, car peu de clients quittent l'entreprise [19], et peu de devis sont annulés.

2.5.2 Méthodes complémentaires

L'analyse de la littérature a révélé plusieurs méthodes hybrides. [7], [13], et [25] utilisent des combinaisons de méthodes différentes pour prévoir et améliorer la prévision générale. Ces articles présentent de différentes manières la complémentarité d'une approche utilisant plusieurs modèles. Les cinq articles réalisent des prévisions par régression, mais y ajoutent une méthode complémentaire. Dans [7], il s'agit de l'utilisation de règles floues (fuzzy rules) pour améliorer les prévisions. [13] utilise un modèle de diffusion en complément. La méthode GARCH présentée dans [25] combine plusieurs modèles de prévision pour diminuer l'erreur finale.

La problématique industrielle discutée dans l'introduction de ce travail est donc peu, voire pas étudiée. L'opportunité de recherche identifiée dans cette analyse de la littérature concerne donc le développement d'un premier modèle de prévision de l'état final de devis de vente. En particulier, ce travail nous permettra de déterminer si une approche utilisant plusieurs méthodes complémentaires permet de bien caractériser les devis et prévoir leur état final.

Tableau 2.5 : Synthèse de la littérature identifiée

Article	Méthode(s)	Devis	Facteurs externes	Secteur	Type de données principal	Type de résultat
[1]	Modèle d'acceptation des commandes	Non	Non	Production	Capacité et importance client	Acceptation de la commande
[2]	Lois de probabilités	Oui	Non	/	Temps de livraison	Temps de livraison devis
[3]	LE, Holtz, ARMAV	Oui	Oui	Électronique	Devis et vente réelle	Prévision de vente
[4]	Forêt aléatoire	Oui	Oui	Vente sur internet	Donnée utilisateur	Prédiction de réalisation
[6]	SVM, ARIMA	Non	Oui	Semi-conducteur	Donnée de vente	Prévision de vente
[7]	Régression linéaire et règles floues	Non	Oui	/	Demande	Prévision de la demande
[8]	RNN	Non	Non	Semi-conducteur	Donnée de vente par produit	Prévision de la demande
[9]	RGM	Non	Oui	Semi-conducteur	Demande	Prévision de la demande
[11]	Régression logistique et arbre de décision	Non	Non	Telecom	Donnée Clients	Prévision de désabonnement client
[12]	KNN, Forêt aléatoire, XGBoost	Non	Non	Telecom	Donnée Clients	Prévision de désabonnement client
[13]	Bass modèle, ARIMA, DARIMA, Lotka-Volterra model vector autoregression	Non	Oui	Semi-conducteur	Donnée de vente et donnée client	Prévision de la demande
[15]	SVM, RNN	Non	Non	Telecom	Donnée Clients	Prévision de désabonnement client
[16]	Arbre de décision et RNN	Non	Oui	Semi-conducteur	Donnée de vente	Classification du temps de production
[17]	Modèle par diffusion	Non	Oui	Semi-conducteur	Donnée de vente	Prévision de la demande
[18]	Naïve bayes, arbre de décision, et la forêt aléatoire	Non	Oui	Telecom	Donnée employés	Prévision de démission
[19]	Arbre de décision, forêt aléatoire, XGBoost	Non	Oui	Telecom	Donnée Clients	Prévision de désabonnement client
[20]	Modèle de tendance aléatoire saisonnière	Non	Oui	Semi-conducteur	Donnée de vente	Prévision de la demande
[24]	ARIMA, Forêt aléatoire et mélange	Non	Oui	Semi-conducteur	Donnée de vente et externes	Prévision de la demande
[25]	GARCH model	Non	Non	Semi-conducteur	Donnée de vente	Prévision de la demande
[28]	RGM	Non	Oui	Semi-conducteur	Donnée de vente	Prévision de la demande
[30]	ARIMA et VAR	Non	Oui	Semi-conducteur	Donnée de vente	Prévision de la demande
[31]	LBVAR	Non	Oui	Industries technologiques	Donnée de vente	Prévision de la demande
[32]	Modèle de diffusion multigénérationnel	Non	Oui	Semi-conducteur	Donnée de vente	Prévision de la demande

(LE : Lissage exponentiel, ARIMA : Moyenne mobile autorégressive intégrée, DARIMA : Moyenne mobile autorégressive intégrée dynamique, SVM : Support vector machines, RNN : Réseaux de neurones récurrents, RGM : Modèle gris roulant, KNN : méthode des k plus proches voisins, XGBoost : Arbre de machine à gradient boosté et Extreme Gradient Boosting, GARCH : modèle vectoriel autorégressif conditionnel généralisé hétéroscédasticité, VAR : modèle d'autorégression vectorielle, LBVAR : modèle d'autorégression vectorielle bayésienne de Litterman

CHAPITRE 3 MÉTHODOLOGIE ET OBJECTIFS

Ce chapitre présente les objectifs du projet de recherche ainsi que la méthodologie générale de recherche suivie.

3.1 Objectif

Le chapitre 1 de ce mémoire nous a permis de contextualiser et clarifier dans un premier temps la problématique industrielle, et plus particulièrement le concept de devis tel qu'utilisé dans l'entreprise étudiée pour affiner manuellement les prévisions de la demande. Le chapitre 2 a mis en avant un manque dans la littérature concernant l'utilisation systématique d'outils avancés d'analyse des données contenue dans les devis afin d'aider le processus de prévision de la demande. Les objectifs de recherche présentés dans cette section découlent directement de cette opportunité de recherche.

Tel que nous l'avons discuté dans le Chapitre 1, les données contenues dans les devis, et plus particulièrement l'évolution dans le temps des données de chaque devis, reflètent la dynamique de la négociation entre le client et le vendeur. Ces données pourraient donc potentiellement être utiles pour « caractériser » le comportement des clients, et ainsi anticiper le devenir de chacun des devis. Puisque les devis sont négociés plusieurs mois avant la réalisation de la vente (ou l'annulation du devis), une telle anticipation permettrait ainsi d'estimer les volumes et le mixte de produits qui seront vendus à une date précise puisque ces derniers contiennent aussi une date de livraison.

Notre question de recherche est donc la suivante :

Peut-on utiliser les données de suivi de l'état des devis pour anticiper leur état final ?

Dans la cadre de ce projet, nous allons donc chercher à répondre à cette question en utilisant les données spécifiques de notre partenaire. Autrement dit, cette étude est limitée par la qualité des données produites actuellement par les processus de gestion des devis en place dans l'entreprise. Nous évaluerons ainsi si ces processus génèrent des données utilisables pour prévoir efficacement l'état final des devis.

L'objectif général en lien avec notre question de recherche est le suivant :

Développer un outil de prévision de l'état final de chaque devis individuellement.

Afin d'atteindre cet objectif, les trois sous-objectifs spécifiques suivants sont proposés :

- *Étudier et décrire les données de suivi des devis ;*
- *Modéliser le cycle de vie des devis ;*
- *Proposer et comparer des modèles de prévision de l'état final des devis.*

3.2 Méthodologie

Le processus général de recherche suivi dans cette thèse est représenté à la Figure 3.1. Ce dernier comporte 9 étapes décrites ci-dessous.

Définition de l'objectif général de recherche : Cette étape a pour but de comprendre dans un premier temps la problématique industrielle, et d'identifier à l'aide d'une analyse de la littérature des opportunités de recherche susceptibles d'apporter une solution novatrice au problème industriel identifié. De nombreuses discussions avec l'entreprise ont été nécessaires afin de comprendre les processus impliqués et les opportunités d'améliorations possibles. Cette étape a ainsi permis de formaliser l'objectif de recherche décrit ci-dessus.

Étude des processus et données industriels : À la suite de l'identification d'une opportunité d'amélioration, des analyses des processus et des données industriels ont été réalisées dans le but de comprendre les caractéristiques spécifiques du problème étudié (processus de vente avec devis), et les données disponibles pour réaliser le projet de recherche.

Analyse préliminaire et ajustements des données : Cette étape a permis de caractériser de façon générale le processus de vente par devis, mais aussi les clients. Cette étape a aussi permis de mettre en lumière les problématiques de la base de données et ainsi lui apporter des corrections. Différentes sélections de données ont ainsi été réalisées pour permettre d'avoir un ensemble de données utilisable. Par exemple, nous avons procédé à la suppression de certaines données inutilisables ou à la correction d'erreurs dues au processus d'enregistrement. À la fin de cette étape, la base de données utile est bien définie. Une analyse descriptive des données a pu être réalisée. Un ensemble de variables indépendantes potentielles pouvant contribuer à prédire le devenir des devis a été identifié.

Définition préliminaire de la fonction de prévision et choix préliminaire des outils : Après l'analyse préliminaire, cette phase a permis de définir la nature de la fonction de prévision.

Puisqu'il ne s'agit pas d'un processus classique de prévision utilisant des séries temporelles, il a fallu définir ce que cette fonction de prévision devait anticiper. Autrement dit, il a fallu développer et proposer des variables dépendantes susceptibles d'être calculées par un outil.

C'est par exemple à cette étape qu'il a été proposé de classifier les devis selon leur état final. De plus, puisque l'analyse préliminaire a mis en avant un certain déséquilibre de la base de données (une grande partie des devis est vendue), il a donc été nécessaire choisir des outils d'apprentissage qui ne soient pas affectés par cette problématique. Tel que proposé par [19], il a donc été décidé de développer les modèles par arbres de décisions et par forêts aléatoires.

Préparation préliminaire des données : À la suite du choix des modèles, cette étape a permis de préparer les données en vue de leur utilisation. Puisque l'objectif de recherche est de prévoir l'état final de chaque devis, une sélection des clients a été nécessaire. Il n'était en effet pas envisageable de développer et tester des modèles de prédictions pour l'ensemble des clients. Les clients ont été sélectionnés selon la quantité et la qualité des données ainsi que par leur importance pour le partenaire. L'encodage des variables a ensuite consisté à déterminer les différentes variables considérées utiles et importantes, et à évaluer leur niveau d'indépendance. Les devis ont deux types de données principales utilisables : les données concernant les produits et les données représentant l'évolution des négociations. L'évaluation de la dépendance des devis entre eux a été un point d'attention durant l'encodage. L'encodage des données et la création de la base ont été faits avec le logiciel Alteryx. Ce dernier est déjà utilisé par l'entreprise et permet donc de faciliter le travail dans les bases de données de l'entreprise. La préparation des données a donc permis la sélection de clients pertinents à la recherche, le calcul de nouvelles variables, et enfin leur encodage pour être utilisable par les modèles. Deux modes de classifications des données ont ainsi été choisis lors de la création des bases de données finales. La première classification « classification simple » est utilisée pour déterminer si le devis est annulé ou vendu. La deuxième classification « classification détaillée » est utilisée pour connaître le quart de vente du devis en plus de l'indication vente ou annulation.

Développement des modèles : L'étape de développement des modèles a intégré leur conception, leur programmation et leur débogage. Il s'est effectué en R avec les bibliothèques Rpart et RandomForest.

Expériences préliminaires et "optimisation" des paramètres des modèles : Cette étape a permis d'améliorer les modèles. Notamment, le choix des indicateurs a été fait à l'aide de la littérature [18]. Les indicateurs choisis ont été la précision, la sensibilité, la spécificité, la justesse, et le score F1. La précision permet de voir d'évaluer le taux de faux positifs. La sensibilité permet d'évaluer les faux négatifs. Dans le contexte d'un test de dépistage d'une maladie, la spécificité du test désigne la proportion d'individus ayant reçu un résultat négatif à ce test parmi ceux qui ne sont pas réellement négatifs. La justesse est la proportion de prédictions correctes (vrais positifs et vrais négatifs) parmi le nombre total de cas examinés. Le score F1 est la moyenne harmonique de la précision et de la sensibilité. Il permet d'évaluer les deux indicateurs en même temps. Ces 5 indicateurs calculés dans les différents cas permettent d'évaluer et de comparer les modèles (méthode de calcul : voir annexe A). Ces indicateurs ont été calculés par deux méthodes. La première est une évaluation qui considère tous les devis égaux. La seconde méthode prend en compte le poids en dollars de chaque devis. Cette seconde méthode a pour avantage de montrer si la classification classe correctement les devis qui ont une grande importance financière. Ensuite, l'optimisation des modèles a été réalisée de plusieurs manières. Pour les arbres de décisions, une analyse par élagage a été faite. Pour les forêts aléatoires, une analyse selon le nombre d'arbres a permis une optimisation des résultats. La validation des résultats a été faite par validation croisée de type *k-Fold*, avec une valeur de k égale à 10. Autrement dit, il s'agit d'une validation croisée avec 90 % des données utilisées pour la base d'entraînement, et 10 % pour la base de test choisis initialement de manière aléatoire. Les résultats des tests ont été analysés client par client pour étudier la fiabilité des modèles un client à la fois. Finalement, cette étape d'expérimentation préliminaire a permis d'orienter les choix de conception des modèles afin de proposer ultimement, et dans le cadre de ce projet, des modèles efficaces. Ainsi, lorsque les résultats préliminaires n'étaient pas satisfaisants, des ajustements ont été proposés. Par exemple, il a été proposé de classifier les clients en fonction de leur comportement d'achat, et d'agréger les données des clients similaires afin d'augmenter la quantité de données disponibles pour fin d'apprentissage des modèles.

Expériences : Une fois les choix modèles programmés, testés et optimisés, un plan d'expérience a été défini afin d'étudier l'impact de différents niveaux d'agrégation des données des clients sur la qualité des modèles proposés. Différentes classifications et agrégations des clients ont donc été réalisées et comparées. Les agrégations des clients ont été faites en utilisant un des trois critères

proposés et testés. Ces derniers sont la hiérarchie de l'entreprise, le poids du client (en $ dépensé), et le mix de produits achetés. Les méthodes de calcul de distance « euclidienne » et « manhattan » ont été évaluées dans cette étude. Après la création des différentes bases de données selon les agrégations des clients et du plan d'expérience, les résultats des expériences ont été obtenus par validation croisée réalisée comme à l'étape précédente.

Analyse des résultats : À cette étape, les résultats ont été analysés selon les modèles et les bases d'entraînements choisis. La classe « devis annulé » a servi de base pour comparer les modèles. En effet, cette classe est la plus pertinente pour l'industriel qui souhaite avant tout savoir si le devis sera vendu. Dans cette partie, l'analyse des variables est réalisée pour comprendre les facteurs importants dans la prise de décisions des modèles. Après cette analyse, les limites et améliorations ont été proposées ainsi que des recommandations pour l'entreprise.

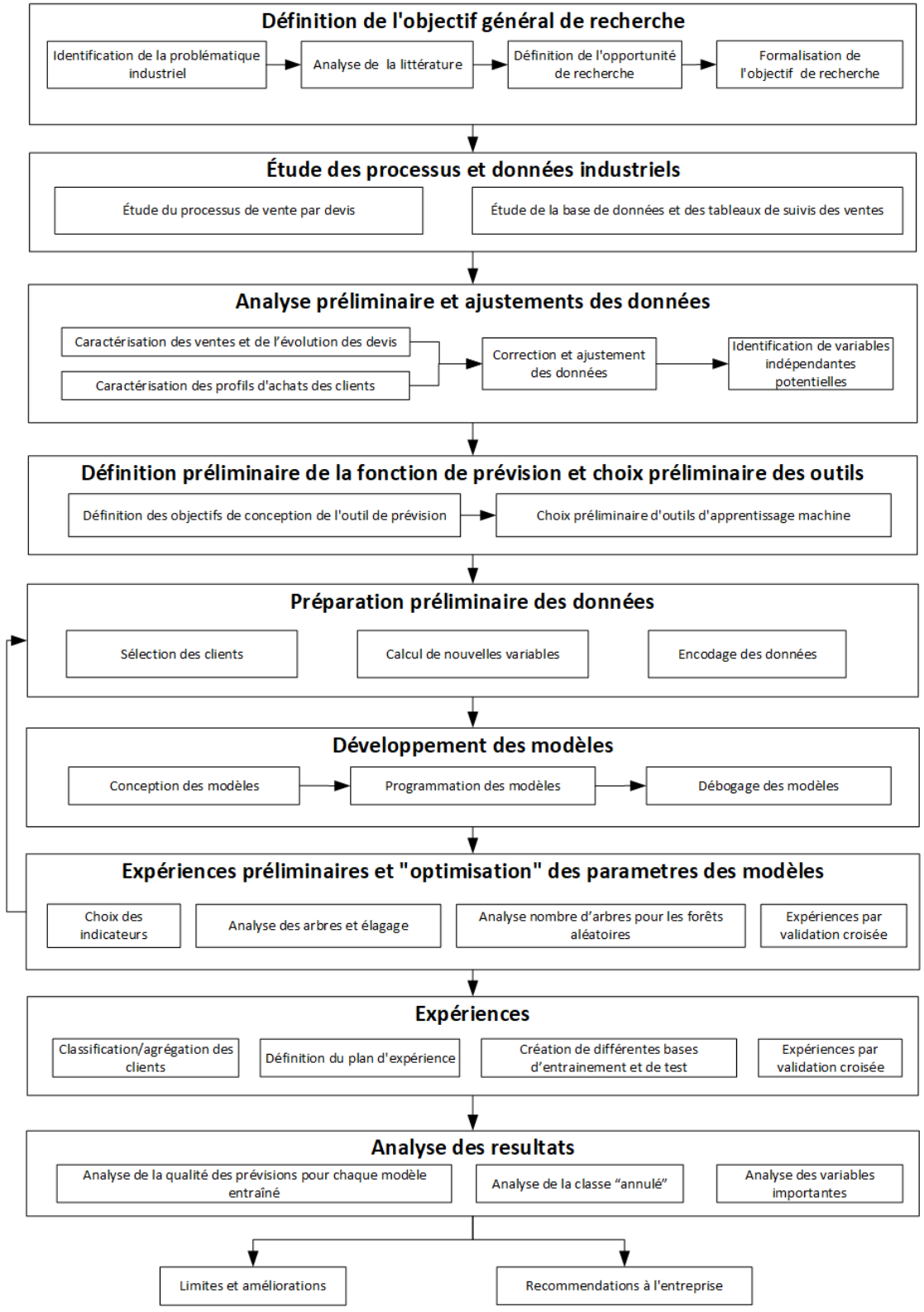

Figure 3.1 : Schéma du processus méthodologique

CHAPITRE 4 PRÉSENTATION ET ANALYSE DESCRIPTIVE DES DONNÉES

4.1 Données disponibles

Les données reçues du partenaire sont décomposées en deux jeux de données.

1. **Un jeu de données représentant l'historique des devis de novembre 2018 à mars 2020 ;**
2. **Un jeu de données décrivant les familles des produits (liste des produits).**

L'entreprise partenaire a donc fourni deux bases de données complémentaires pour la recherche. La première base de données correspond à l'historique des devis. Elle permet de voir l'évolution des devis dans le temps. La deuxième base de données permet de relier un numéro de produit à ses informations descriptives, telles que sa gamme, son prix ou son importance au sein de l'entreprise par exemple.

Dans la suite de ce chapitre, les colonnes des bases de données seront présentées puis une analyse descriptive de ces données sera proposée.

4.1.1 Base de données Devis

Ce jeu de données présente 24 colonnes pour environ 20 millions de lignes. Les données correspondent à la liste des devis de l'entreprise sur la période novembre 2018 à mars 2020 avec une vue des devis en cours. Chaque semaine, une capture d'un instantané des devis est prise. Le fichier permet de suivre l'évolution des devis dans le temps, semaine après semaine. La base est donc le regroupement des « photos » des devis.

4.1.1.1 Colonnes « Temps »

Les colonnes 1 et 2 contiennent les informations qui permettent de suivre la temporalité de la base et des devis. La première colonne donne la date de la « photo » et la deuxième la date de création du devis.

1- Date photo (Date Stamp) - *Date* :

Il s'agit de la date à laquelle l'instantané est pris. Cette colonne permet de suivre l'évolution du devis à des dates précises.

2- Date de création - *Date* :

Il s'agit de la date de création de devis.

4.1.1.2 Colonnes « État client-vendeur »

Les colonnes 3 à 6 renseignent sur l'état actuel du devis dans le processus de négociation de l'entreprise entre le client et le vendeur. Ces données peuvent varier dans le temps.

3- Probabilité du devis - *Entier* :

Il s'agit de la probabilité, estimée par le vendeur, que le devis se matérialise en une commande ferme. Le vendeur doit donner cet indicateur de confiance une valeur entre 0 et 100 %. À partir de 80 %, les devis sont pris en compte par le département de prévision, il s'agit du filtre vu dans la partie 1.2.2.2. Ce nombre est donc un indicateur subjectif non calculé. Il n'existe pas chez le partenaire de méthode standard d'estimation de ce processus d'attribution.

4- Données DC consolidé (DDCC) – *Chaîne de caractères* :

La colonne « données DC consolidé » indique l'état du devis dans son processus logistique. Il permet de savoir si le devis est en cours de réalisation, « VENDU » ou expédiée.

L'entreprise a séparé les devis en cours de réalisation en trois sous-catégories selon la probabilité donnée par le vendeur. Cela permet un tri rapide entre les devis qui ont une forte confiance de réalisation des autres.

L'état « SIEBEL 0 % TO 50 % » correspond à un devis pour lequel la probabilité de réalisation est comprise entre 0 et 50 %.

L'état « SIEBEL 51% TO 79% » correspond à un devis pour lequel la probabilité de réalisation est comprise entre 51 et 79 %.

L'état « PREVISION » correspond à un devis pour lequel la probabilité de réalisation est supérieure à 80 %.

5- Stade du devis – *Chaîne de caractères* :

La colonne stade du devis permet de suivre l'état de négociation du devis et notamment si le devis a été annulé. Nous avons identifié deux types de stades. Les stades de négociations comme « Solution en développement » ou « Proposition » qui permettent de suivre l'évolution du processus. Les stades de fermetures de devis par annulation comme « Fermé – Annulé » ou « Fermé Perdu ». La liste complète des stades et de leur signification est visible dans le Tableau 4.1.

Tableau 4.1 : Liste des stades

Valeur du stade	Signification
Fermé - Annulé (reporté)	Annulation
Fermé - Créé par erreur	Annulation
Fermé Perdu	Annulation
Fermé Gagné	Vendu
Fermé – en retard	Annulation
Solution en développement	Négociation en cours
Inactif	Inactif
Négociation	Négociation en cours
Propose	Négociation en cours
Prospection et identification des opportunités	Négociation en cours
Qualification	Négociation en cours
Vide	Vendu

6- Date souhaitée par le client (CRD) – Chaîne de caractères

Il s'agit de la période fiscale de livraison demandée par le client. Dans l'entreprise partenaire, l'année est découpée en quatre périodes fiscales. Chaque période fiscale est elle-même divisée en 3. Ces colonnes permettent d'identifier le mois de livraison demandé par le client. Par exemple, si le client veut être livré en mars 2019, ces colonnes indiqueront FY19Q2 en période fiscale et 2 en mois fiscal.

4.1.1.3 Colonnes « Produits »

Les colonnes 7 à 10 renseignent sur le contenu du devis telles que le type de produit, et leur quantité.

7- Gamme produit (ITEM Load Plan Product Line) – *Chaîne de caractère :*

Il s'agit de la gamme du produit. C'est le regroupement le plus large.

8- Numéro produit (Item Number) – *Chaîne de caractère* :

Il s'agit du numéro d'identification du produit. C'est le plus petit niveau de classification produit.

9- Quantité de produit – *Entier* :

Il s'agit de la quantité demandée par le client pour un produit dans le devis et/ou la commande.

10- Valeur d'inventaire en $ - *Flottant :*

Cette colonne indique la valeur avant marge de la ligne. (= Coût standard du produit *Quantités)

4.1.1.4 Colonnes « Informations statiques »

Les colonnes 11 à 16 donnent les informations statiques du devis, telles que son nom et son client.

11- ID Devis (Quote) – *Chaîne de caractère* :

Cette colonne indique l'identifiant unique du devis. (Exemple de ID Devis : A0E7MX3X)

12- Numéro de commande – *Chaîne de caractère* :

Il s'agit de l'identifiant de la commande à la suite d'un devis. Un devis peut être vendu sous différents numéros de commande (1.2.2.1).

13- Région ID - Région de planification - DP Sub Cust Top Client - Planification Top Client

Il s'agit de l'identification des clients et de la subdivision. Les régions ont pu changer par le passé. Il y a en tout un total de 614 clients.

Exemple : Client X :

- Région ID : 5
- Région de planification : 6
- DP Sub Cust Top Client: 96
- Planification Top Client : 8
- Numéro Client : 24554

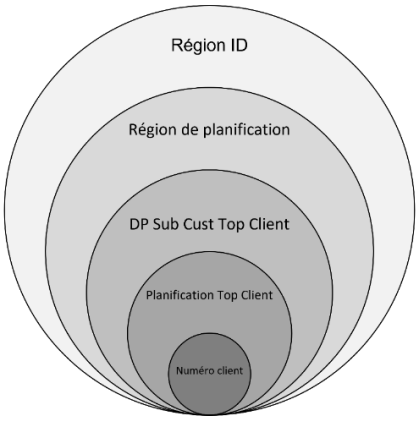

Figure 4.1 : Hiérarchie client

4.1.2 Base de données Produits

Ce jeu de données permet à partir du numéro du produit de récupérer toutes ses informations. On peut par exemple retrouver la hiérarchie de sa famille de produits, son coût ou son importance. Ce fichier permet de gérer les produits en les agrégeant plus ou moins en suivant la hiérarchie des produits.

Les produits sont ainsi hiérarchisés en 4 strates. Dans les données, on part de 11 *gammes de produits*, puis 233 *Sub KFU code*, 1393 *familles de produits* et enfin 1627 *numéros produits* (Figure 4.2).

En plus de cette hiérarchie, les produits sont classés en 3 catégories selon leur importance.

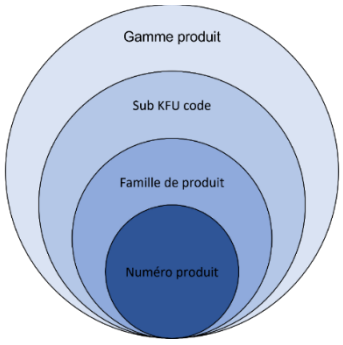

Figure 4.2 : Hiérarchie produit

La colonne *Portfolio Driver* gère cette importance. Les trois catégories sont les suivantes : *High driver*, *Medium Driver* et *Low Driver*. Les produits les plus importants sont regroupés sous l'appellation « High Driver ».

4.2 Analyse descriptive des données et des processus

Nous allons présenter dans cette section une analyse approfondie des données. Cette dernière a permis d'apprendre de nombreuses informations sur les devis, les clients et les processus, mais également de mieux comprendre la nature des données et les défis du problème d'analyse des données.

4.2.1 Les types de devis

Dans l'entreprise partenaire, il existe deux types de devis ou processus de vente. Les deux processus cohabitent. Certains clients peuvent avoir recours exclusivement à l'un des deux ou utilisent un mélange des deux. Les deux processus sont les suivants :

- Le processus de vente classique avec négociation, que nous appellerons « devis classique ».
- Le processus en vente directe, les « devis en vrac » (bulk quote) ;

Les devis en vente directe ou « bulk quotes » ne fonctionnent pas sur le principe de vente par devis et doivent être analysés différemment. En effet, ils sont directement vendus aux clients sans délai. Ils ne peuvent donc pas être prédits à partir d'une analyse du cycle de vie des devis.

Les sections suivantes décrivent les états possibles des cycles de vie des deux types de devis.

1. Devis classique (cas 1 et 2 de la Figure 4.3)

Ce type de devis a un fonctionnement relativement simple dans une majorité de cas. La Figure 4.3 présente les différents cycles de vie possibles d'un devis classique.

a. Création du devis

À sa création, le devis reçoit plusieurs informations. Certaines sont non modifiables telles que l'ID du devis ou son client. Tandis que d'autres sont assujetties à modification telles que la liste des produits ou la probabilité de réalisation donnée par le vendeur.

b. Négociation du devis

Pendant la négociation du devis, ce dernier évolue. Les produits et les quantités peuvent changer, mais aussi la probabilité de réalisation, le stade du devis, etc. Ces changements sont liés à

l'évolution des besoins des clients. Le CRD peut aussi changer selon les disponibilités du partenaire et les délais du client.

c. Fin du devis

À la fin de son cycle de vie, le devis aura été réalisé ou annulé. Sa réalisation sera marquée par la création d'un numéro de commande et le changement du DDCC en « ORDRE DE VENTE ».

L'annulation d'une commande sera quant à elle marquée par un changement de stade du devis vers un des différents stades d'annulation.

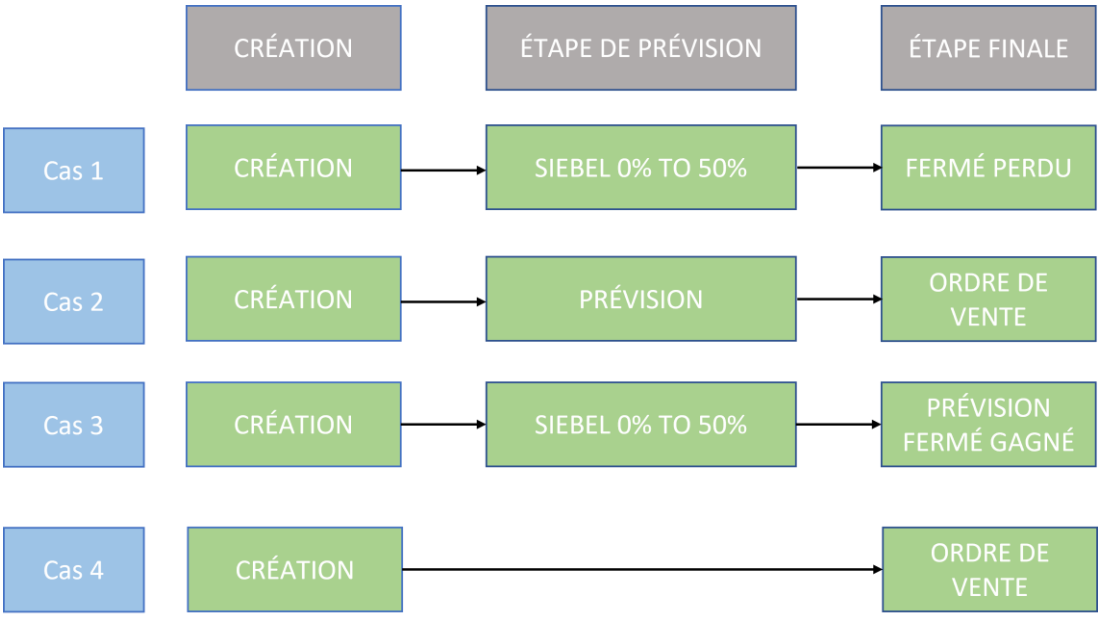

Figure 4.3 : Différents parcours de vie des devis

2. Devis en vrac (cas 4; « bulk quotes » de la Figure 4.3)

Les devis en vrac fonctionnent sur un principe de vente directe. Ces devis sont constamment ouverts. Autrement dit, les produits vendus sont insérés et retirés lorsqu'ils sont expédiés sans que le devis ne disparaisse. Les produits arrivent ainsi directement dans un mode vendu. Ils ne restent que le temps d'être expédiés au client.

Les produits qui passent par ce type de vente n'ont pas de parcours de vie visible dans les données.

3. Les exceptions (cas 3 de la Figure 4.4) :

Certains devis ont un comportement particulier. Ceux-ci ont un parcours de vie sur le principe des « devis classiques », mais ont une fin particulière. Juste avant d'évoluer vers un état de vente, les devis passent dans l'état :

- Stade du devis est « FERME GAGNÉ » ;
- DDCC égale « PREVISION ».

Les devis qui s'arrêtent dans cet état n'évoluent plus. Ces devis sont en fait vendus. Cependant, leur vente finale est enregistrée via un devis en vrac. Ces produits sont donc vendus, mais sous un autre devis, ce qui complexifie significativement leur traitement.

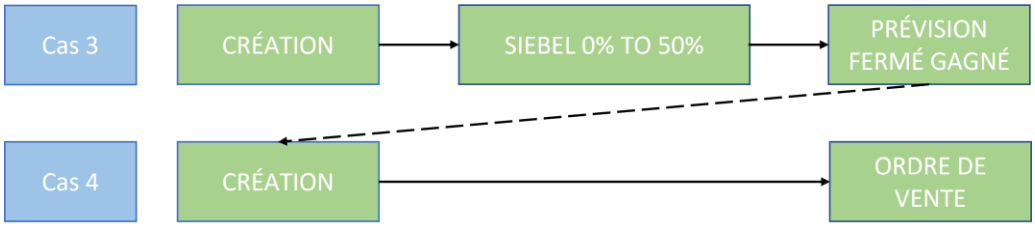

Figure 4.4 : Cas spécial

4.2.2 Cycles de vie des devis

Une des premières analyses nous a permis de comprendre l'évolution des devis dans le temps, ainsi que les comportements d'achat et de négociation des clients.

4.2.2.1 Transitions des devis entre différents états

La Figure 4.5 ci-dessous représente pour un client particulier l'évolution de ses devis dans le temps dans certains états. Chaque carré représente un état dans le processus de vente d'un devis. L'épaisseur des flèches représente la fréquence observée des transitions entre ces états dans la base de données pour ce client (plus le trait est épais entre deux états, plus il y a d'observations de transition entre ces états). Cette représentation permet donc suivre les transitions d'état les plus observées.

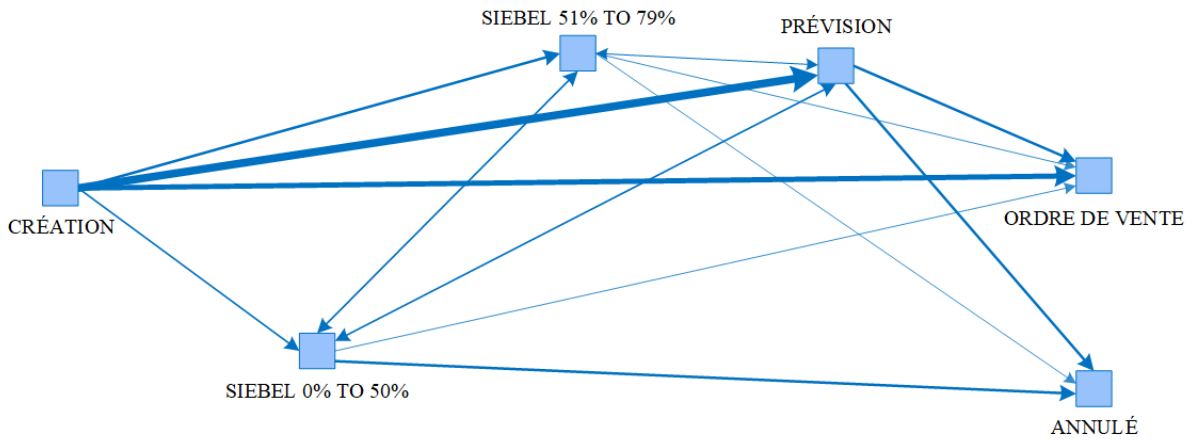

Figure 4.5 : Évolution de la vie du devis (Client 23)

La ligne directe entre la création et l'ordre de vente correspond aux ventes directes. Ces devis sont impossibles à prévoir en utilisant les données de cycle de vie puisqu'il n'y en a pas. En effet, une partie d'entre eux sont des achats sans préavis, alors que d'autres sont des ventes faites en amont avec l'état « PREVISION-FERME GAGNÉ » (Figure 4.3, cas 3).

La Figure 4.5 permet ainsi de mieux comprendre ce que l'on cherche à prédire. Les devis évoluent ainsi plus ou moins rapidement avant d'entrer dans un état de certitude de vente.

Ces graphiques nous ont aussi permis d'identifier des problèmes dans les données : l'évolution impossible allant de l'état « ANNULE » vers « SIEBEL 0 % - 50 % » par exemple. Ces problèmes ont été corrigés lors du nettoyage des données.

4.2.2.2 Temps dans un état

Une autre analyse pertinente concerne le nombre de semaines durant lesquelles un devis reste dans un même état. Cette analyse nous permet notamment de mieux comprendre le type de devis étudié.

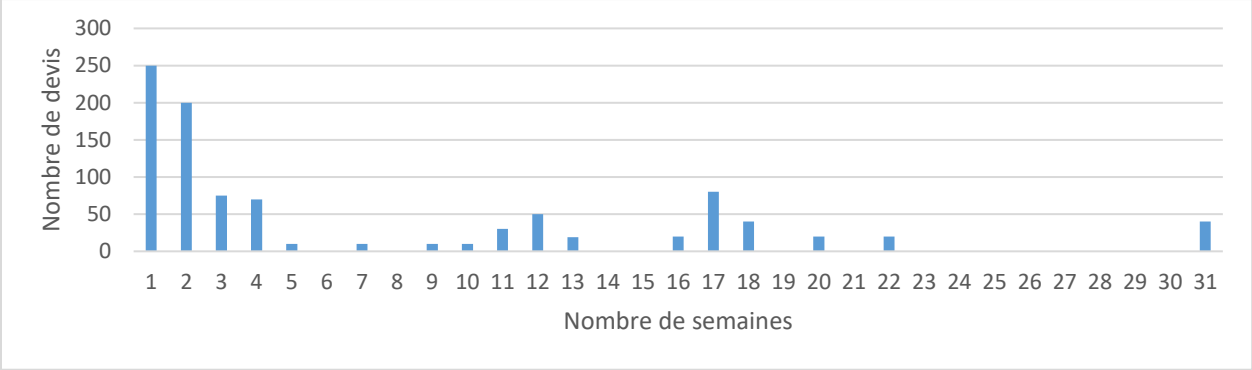

Figure 4.6 : Nombre de semaines durant lesquelles les devis sont restés dans l'état « PREVISION » (en venant de l'état Création) (Client 23)

Par exemple, pour le client 23, nous pouvons voir sur la Figure 4.6, la durée en semaines pendant laquelle les devis restent dans l'état « PREVISION ». Les devis, tous clients confondus, restent entre 1 et 8 semaines dans cet état avant d'évoluer. Les devis restant plus de 8 semaines dans cet état correspondent à des devis que l'entreprise a oublié de supprimer après une vente ou une annulation (Partenaire). La préparation des données doit aussi tenir compte de ces oublis. Pour les « SIEBEL 0 % TO 50 % » et « SIEBEL 51 % TO 79 % », les devis y restent davantage de temps en moyenne. Pour ces deux états les devis y restent entre 1 à 12 semaines en majorité.

4.2.2.3 Âge à l'arrivée dans un état

L'étude du cycle de vie nous a aussi menés à analyser l'âge des devis lorsqu'ils changent d'état. La Figure 4.7 présente, pour la région 1, l'âge en nombre de mois des devis lorsqu'ils arrivent dans les différents états. Par exemple, on remarque que plus de 200 devis ont été créés depuis moins d'un mois lorsqu'ils arrivent dans l'état « SIEBEL 0 % TO 50 % ». Cette figure permet aussi d'observer que les devis arrivent très jeunes dans l'état de « Prévision » et évoluent rapidement dans un état final : dans les 4 mois suivant leur création. Les derniers devis qui ne se trouvent pas dans un état final sont, comme précédemment, des devis qui ont été oubliés. On retrouve ce motif chez la majorité des clients.

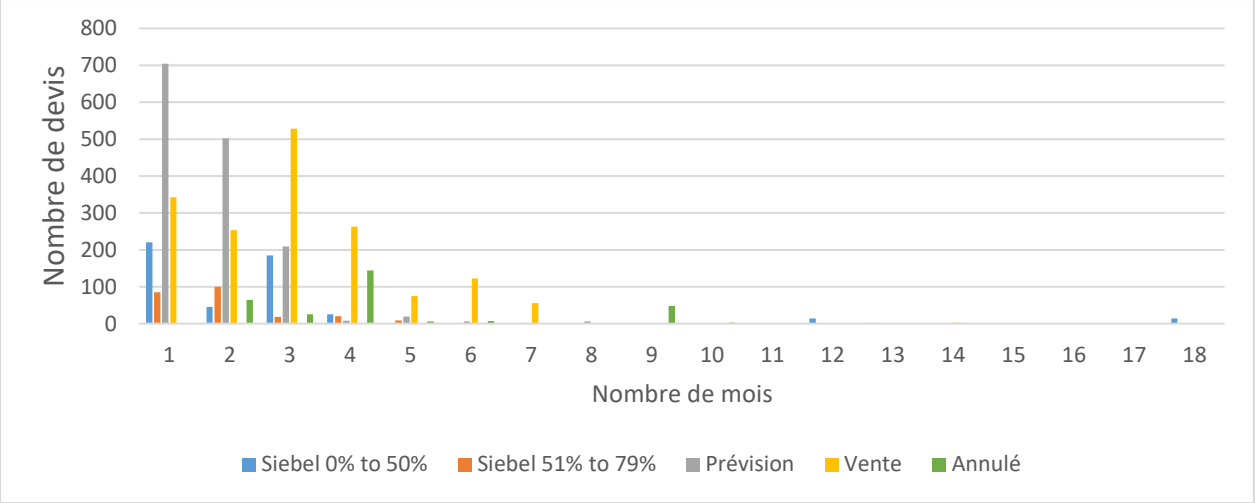

Figure 4.7 : Âges des devis à l'arrivée dans cet état pour la région 1

On remarque aussi que beaucoup de devis jeunes (700) arrivent directement dans l'état « PRÉVISION ». Ces analyses nous ont donc permis de comprendre que l'utilisation des données des cycles de vie des devis ne pourrait uniquement permettre d'améliorer la prédiction de leur état final dans une fourchette de temps entre 1 et 4 mois.

4.2.3 Liens entre les produits

Les clients de l'entreprise partenaire ont des besoins et des comportements distincts. Les produits vendus par l'entreprise répondent à leurs besoins spécifiques. Les paniers de produits varient fortement d'un client à l'autre. Les deux graphiques ci-dessous (Figure 4.8) présentent des exemples pour deux clients. Le client 23 achète fréquemment les familles de produits 3, 5, 6 et 7 ensembles. Le client 125 achète, quant à lui, les familles de produits 5 et 6 ensembles. L'analyse de la composition des paniers de produits nous permet d'identifier les clients qui ont le même comportement d'achat et qui sont donc susceptibles d'avoir un comportement similaire dans leurs achats futurs. Cet aspect est analysé par la suite.

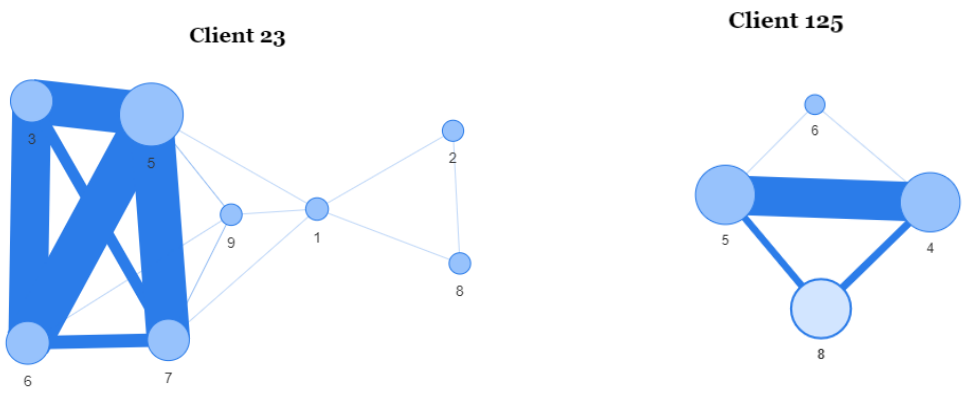

Figure 4.8 : Liens entre les gammes de produits (client 23 et 125) des analyses clients

4.2.3.1 Hétérogénéité des données

L'hétérogénéité des données individuelles des clients ne peut pas mener à une qualité homogène de prévision. Les graphiques ci-dessous (Figure 4.9) représentent l'évolution par mois des prévisions et réalisations des ventes pour deux clients de 90 jours avant le quart jusqu'à la fin de celui-ci. La partie verte représente les parts des ventes cumulées réalisées au cours du quart de vente. La partie bleue représente les ventes annulées. La partie jaune représente la part des ventes encore en cours de négociation, donc prévue, mais non réalisée. Le client présenté sur le graphique de gauche anticipe peu ses commandes futures. Autrement dit, la partie des ventes négociées avant le début du quart ne représente ainsi qu'un tiers des ventes finalement réalisées. Au contraire, le client présenté sur graphique de droite présente un comportement différent. Au début du quart, le client a une bonne visibilité sur ses commandes futures. Cependant, pendant les mois suivants, de nombreux devis sont annulés et de nouvelles prévisions apparaissent. Certaines d'entre elles seront finalement décalées au quart suivant. Ce type de décalage rend le processus de prévision, et donc le lancement des ordres de production, plus difficile. L'anticipation de leur décalage pourrait donc potentiellement avoir un impact positif sur la planification de la logistique.

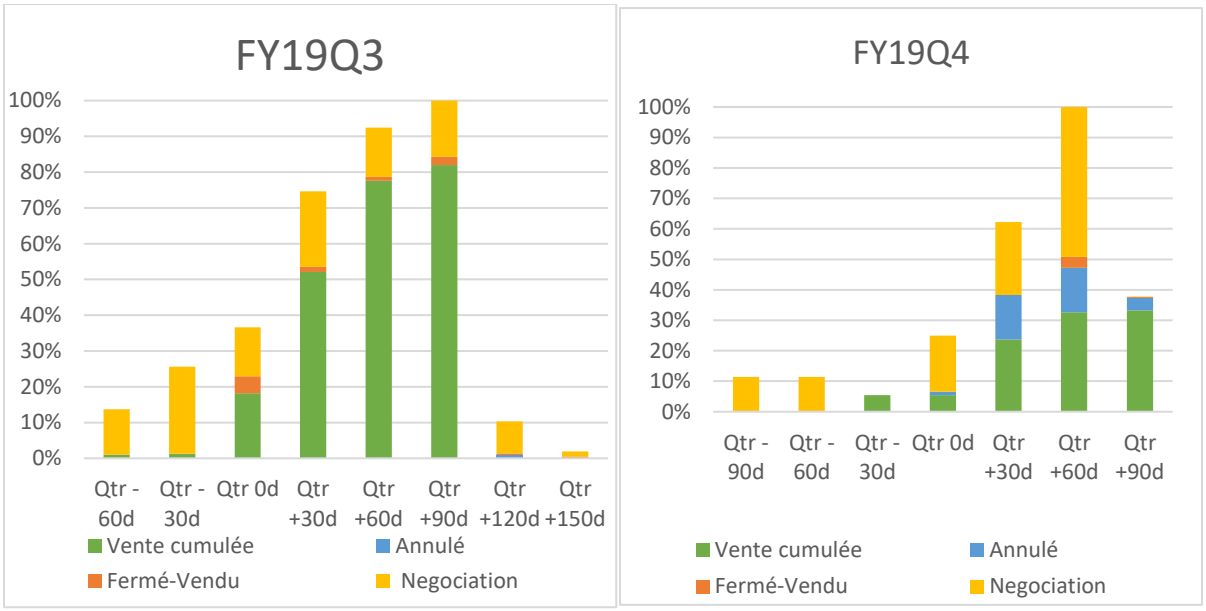

Figure 4.9 : Prévisions et ventes pour deux clients

4.2.3.2 Ventes directes VS Ventes par devis

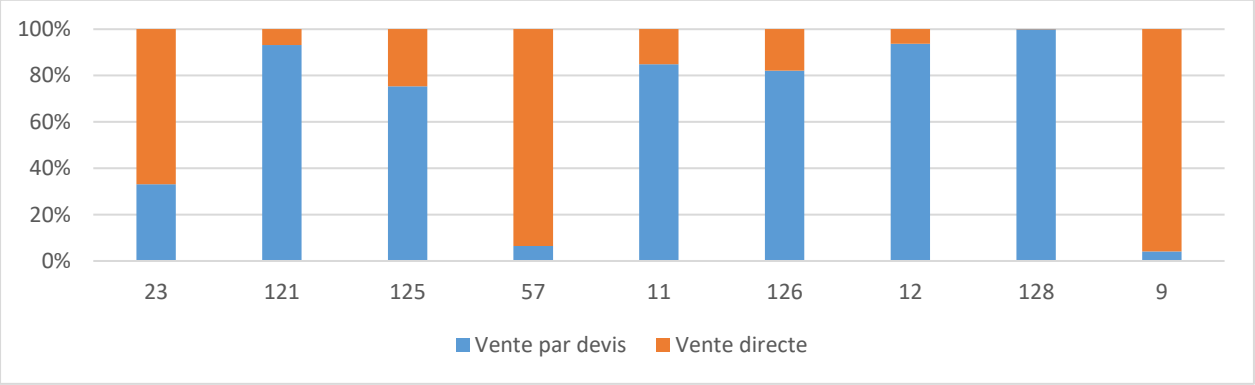

Figure 4.10 : Proportions (en dollars) ventes directes / ventes par devis pour différents clients

La proportion de ventes directes est très variable selon les clients. On peut voir dans la Figure 4.10 que certains clients achètent quasi exclusivement en vente directe (clients 57 et 9) alors que d'autres se concentrent sur les ventes par devis (clients 121 et 128 par exemple). Enfin, il existe un troisième groupe de clients, comme le client 12, qui utilise les deux modes d'achats de manière équivalente.

Le projet présenté dans ce mémoire se concentre sur la prévision des ventes par devis (en bleu).

CHAPITRE 5 EXPÉRIENCES – MODÈLES - RÉSULTATS

Cette partie décrit dans un premier temps la préparation et la réparation des données. Ensuite différentes bases de données seront présentées pour construire les modèles et évaluer leur efficacité en utilisant le processus de validation croisée décrit dans le chapitre précédent. La modélisation et l'optimisation de deux modèles concluront cette partie avec la présentation des résultats détaillés pour un client, et une synthèse des résultats pour 5 clients différents. Les résultats détaillés pour les autres clients sont présentés dans les annexes C à G.

5.1 Préparation des données

5.1.1 Préparation initiale des données

Les données fournies par le partenaire présentent une irrégularité concernant le pas de temps des « photos » des devis. En effet, pour les premières semaines, l'entreprise a pu fournir une « photos » de l'état des devis par jour. Pour que cela ne biaise pas les résultats, nous avons conservé un rythme régulier entre deux « photos ». Une seule « photos » par semaine a donc été conservée pour la suite de l'étude.

5.1.2 Réparation du processus d'enregistrement des données

Avant de procéder à l'analyse des données, nous avons dû procéder à une première réparation des données liées à leur processus d'enregistrement chez le partenaire. En effet, un problème est rapidement apparu. Ainsi, lorsqu'un devis est vendu, il change de statut, et passe du statut « PREVISION » à « ORDRE DE VENTE ». Ce changement d'état permet de valider la vente. Cependant, 20 % des devis analysés évoluent dans un autre mode de vente (cas 4 de la Figure 5.1). Au lieu des séquences présentées dans la section sur le cycle de vie des devis, ces devis suivent la séquence suivante :

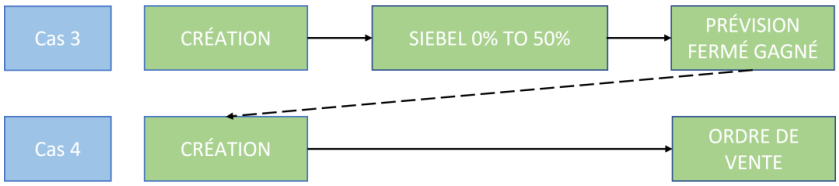

Figure 5.1 : Cas spécial de vente par l'état : « PRÉVISION – FERMÉ GAGNÉ »

L'état « PRÉVISION – FERME GAGNE » devient un état final sans qu'il y ait eu de création de numéros de vente. Le devis reste donc dans cet état jusqu'à sa suppression éventuelle. Cependant, après vérification avec l'entreprise, il se trouve que ce devis a été vendu avec les mêmes produits et les mêmes informations de destination, mais sous un autre devis avec un nom de devis/route différent.

Nous avons donc décidé de résoudre ce problème, car considérer les devis dans cet état comme vendus pose un autre problème puisque les produits inclus dans ces devis apparaissent vendus deux fois.

5.1.2.1 Essais de réparation

La première tentative de résolution de cette problématique consiste à « recoller » ensemble les devis correspondants.

5.1.2.1.1 Approche générale

Pour résoudre ce problème, nous avons donc créé un programme (Figure 5.2) qui sélectionne les devis à comparer afin de savoir s'ils correspondent au même processus spécifique de vente.

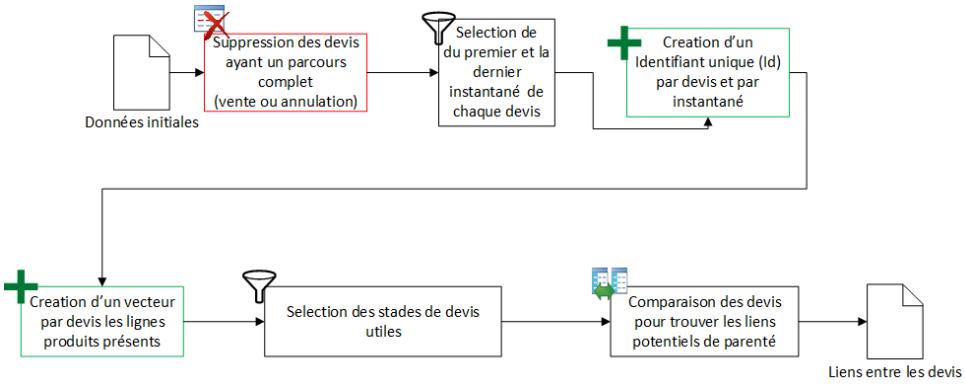

Figure 5.2 : Processus de résolution

Le programme fonctionne selon le principe suivant :

1. Exclusion des devis ayant déjà un parcours complet.
2. Sélection de la première et de la dernière image du devis. Cette étape permet diminuer le nombre de comparaisons. Il suffit en effet de comparer seulement la fin et le début de devis. Les états intermédiaires sont inutiles.
3. Création d'un identifiant unique : devis + date.
4. Exclusion des produits « Low driver ». Ces derniers ne sont pas pertinents et alourdissent le processus. Cette étape facilite la comparaison entre les devis.
5. Création d'un vecteur pour chaque devis pour permettre la comparaison. Dans ce vecteur on retrouve les données suivantes : la date, les produits, le client (Région, Dp sub cust).
6. Sélection des devis qui ont les états « PRÉVISION – FERME GAGNE » ou « ORDRE DE VENTE – *VIDE* ». En effet, il est inutile de comparer des débuts ou des fins entre eux.
7. Comparaison des devis

5.1.2.1.2 Comparaison des devis

La comparaison des devis est visible dans la séquence de la Figure 5.3. Cette comparaison se fait selon plusieurs étapes, et a pour but de faire ressortir les devis similaires. Chaque étape évalue les différences entre deux devis. En particulier, nous avons regardé les critères suivants :

1. Vérification de la logique temporelle. La date de fin d'un devis doit être antérieure à celle du début du suivant.
2. Calcul de la différence de valeur des devis.
3. Calcul de la distance entre les clients selon la classification de l'entreprise partenaire.
4. Calcul de trois distances entre la composition et les délais demandés par les clients.

Distance A = $\sqrt{\sum(NP_{devis1,i} - NP_{devis2,i})^2}$ (1)

Avec $NP_{devis1,i}$, le nombre de produits de la ligne de produit i dans le devis 1.

Distance A_{bis} =$\sqrt{\sum(P_{devis1,i} - P_{devis2,i})^2}$ (2)

Avec $P_{devis1,i}$, la présence (1) ou l'absence (0) de la ligne de produit i dans le devis 1.

Distance B = $TrimestreLivraison_{devis1}$ - $TrimestreLivraison_{devis2}$ (3)

5. Calcul de la distance finale

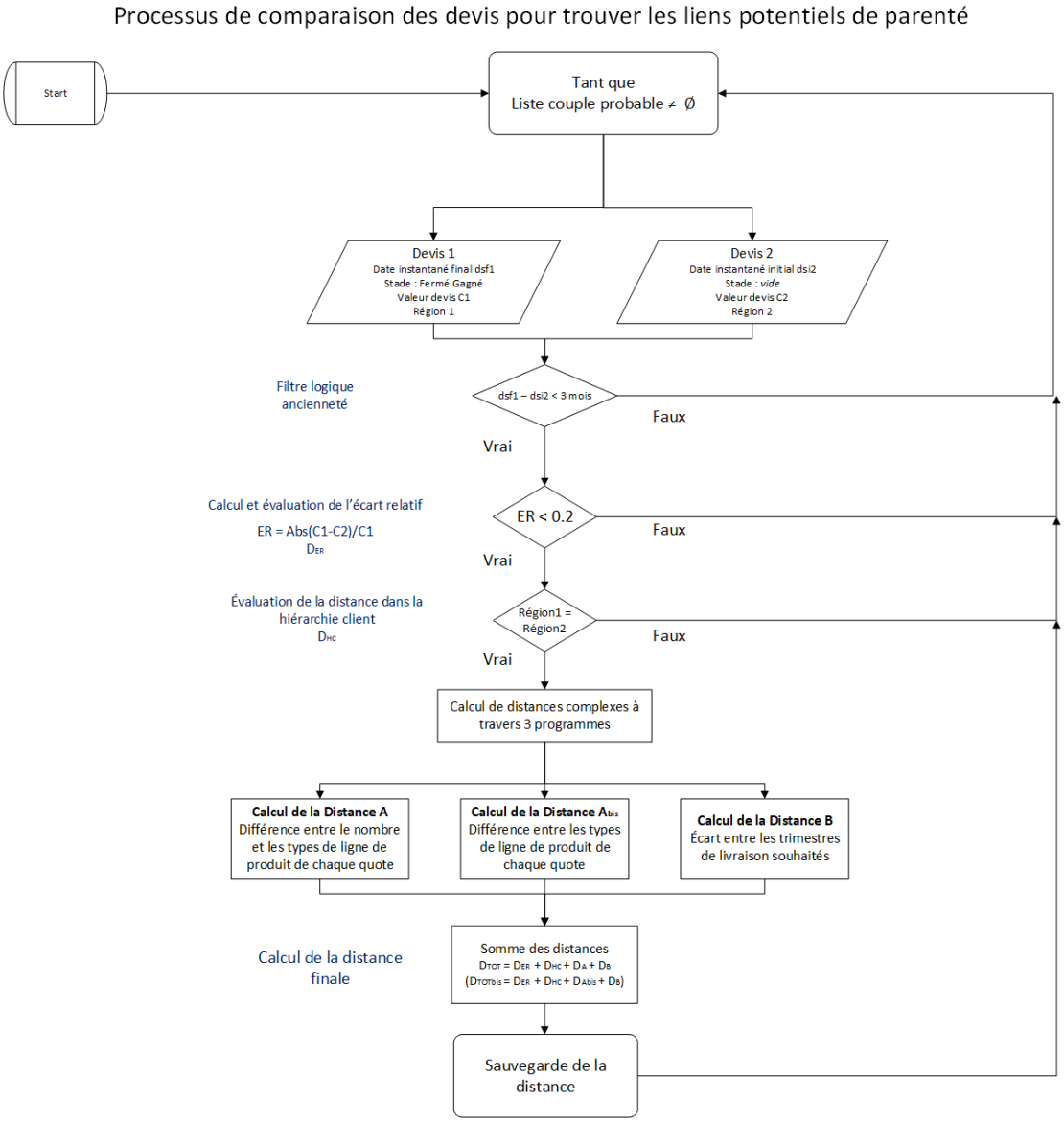

Figure 5.3 : Processus de comparaison

À la suite de la comparaison des devis, on obtient pour chaque devis, une liste de devis similaires possibles et les distances associées. Cependant, comme le processus de saisie des informations par les vendeurs n'est pas standard, il n'a pas été possible de pousser davantage l'automatisation de l'identification des devis similaires. Par conséquent, malgré cette automatisation, les listes des

devis similaires à vérifier manuellement étaient trop importantes. Si certaines réparations sont facilement réalisables par exemple, car les noms de route sont similaires. L'analyse de chaque devis prend ainsi entre 5 et 10 minutes. Comme la base de données contient des milliers de devis, la réparation complète de la base de données prendrait beaucoup plus de temps que celui alloué à cette étude.

Une autre difficulté est qu'un devis qui entre dans l'état « PRÉVISION – FERME GAGNE » n'évolue pas forcément en un seul devis de vente. Il arrive qu'un devis engendre plus d'une dizaine de sous-devis de vente. Dans ce cas, il devient impossible de reconnecter les devis ensemble et de réécrire une continuité logique entre les différentes dates.

Enfin, les clients peuvent changer pendant l'évolution du devis à cause de l'évolution de la classification des clients dans l'entreprise. Comme ces changements ne sont pas consignés systématiquement par l'entreprise, il est impossible d'utiliser cette méthode pour réparer la base de données initiale. Afin de résoudre cette problématique, et éliminer les devis similaires de la base de données, une solution alternative a été mise en œuvre.

5.1.2.2 Solution alternative à la réparation

Le principe général mis en œuvre pour réparer la base de données initiales est de restreindre l'étude du cycle de vie aux états pertinents pour prévoir l'état final du devis. Ce processus est décrit ci-dessous.

5.1.2.2.1 Réparation

Pour résoudre ce problème de « recollage » des devis, nous avons proposé une solution qui consiste à dire qu'un devis qui arrive dans l'état « PRÉVISION – FERME GAGNE » est vendu à la date de première apparition de cet état. On transforme alors l'état « PRÉVISION – FERME GAGNE » en état « ORDRE DE VENTE – « VIDE » ». Bien que cette solution soit simple et rapide, elle comporte un inconvénient majeur qui est l'apparition de doublons des ventes dans le fichier final. Cependant, ce problème n'est pas contraignant pour la suite, car nous cherchons à prédire la fin du devis. Les devis en « VENTE DIRECTE » ne pouvant pas être prédits par une analyse de cycle de vie, ils sont automatiquement supprimés lors des entraînements des modèles.

5.1.2.3 Prise en compte des répétitions

En plus de la nécessité d'ajuster les devis dont le cycle de vie est artificiellement tronqué par le processus administratif et de gestion des ventes, une seconde réparation a aussi été nécessaire avant de procéder à l'analyse des données. Cette réparation concerne les répétitions de devis.

Ces répétitions sont différentes des doublons mentionnés dans la section précédente. Elles sont la duplication de devis identiques : même cycle de vie, même client final, mêmes produits achetés, et en quantité égale. Cependant, ces répétitions ne sont pas des doublons. Ces devis n'ont pas été créés par erreur. Ils sont en fait la subdivision d'un grand devis en plusieurs petits devis égaux. Cet artéfact est comme précédemment issu du processus administratif et de gestion des ventes du partenaire. Afin d'éviter le surentraînement qui pourrait apparaître à la suite de l'utilisation de ces devis interreliés, ils ont été regroupés. Une colonne faisant état du nombre de répétitions est créée afin de conserver cette information et savoir si la subdivision impacte l'état final du devis.

Pour cela, nous avons écrit un programme (Figure 5.4) pour identifier ces devis « répétés ». Une fois les données nettoyées et filtrées, la comparaison entre les devis se fait en utilisant la date de création, la route, les produits à l'état initial, le client et la valeur initiale du devis.

Une nouvelle colonne « répétition » a ainsi été créée. Elle indique le nombre de devis présentant les mêmes critères. Les devis identiques sont ensuite regroupés avec le même nom.

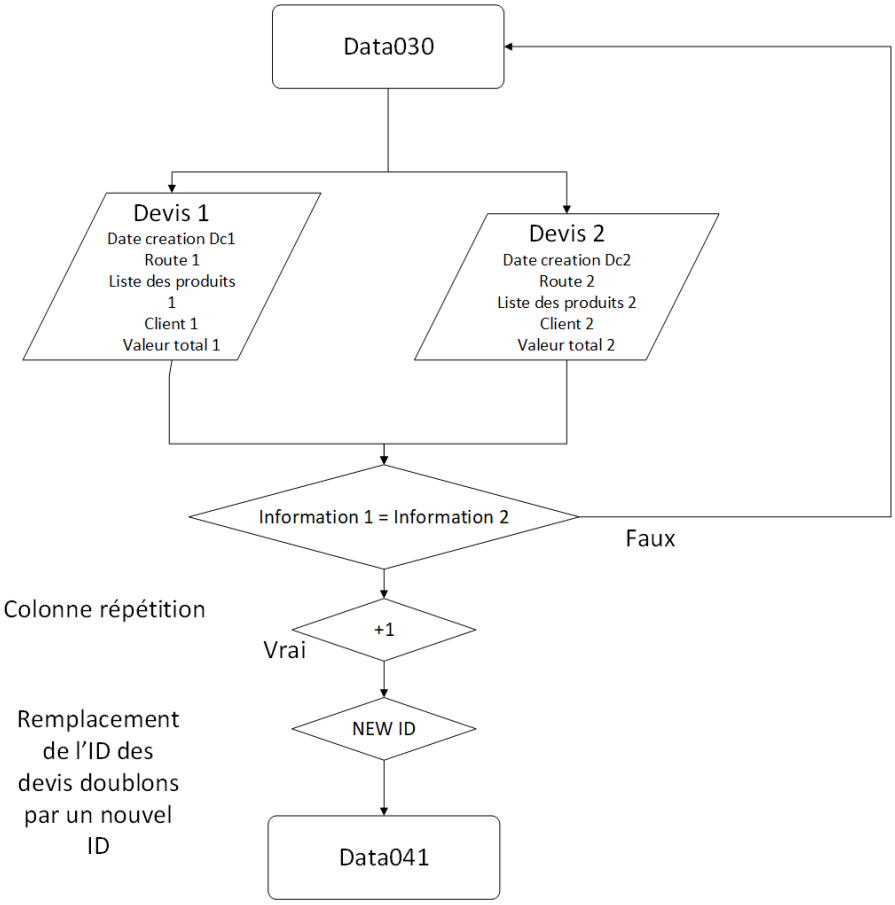

Figure 5.4 : Processus de comparaison des devis

5.1.3 Sélection des devis utilisables

Une fois les réparations faites, comme le partenaire nous a fourni des données de l'état des devis entre le 6 novembre 2018 et le 29 avril 2020, il a fallu une nouvelle fois appliquer un filtre pour être certain d'avoir des devis complets à analyser. Autrement dit, nous avons dû retenir ceux qui avaient un cycle de vie complet entre ces deux dates. Les devis n'ayant pas de date de création ou de finalisation ont donc été exclus de l'étude. Les Figure 5.5 et Figure 5.6 présentent respectivement les processus de nettoyage et de sélection des devis utilisés ici.

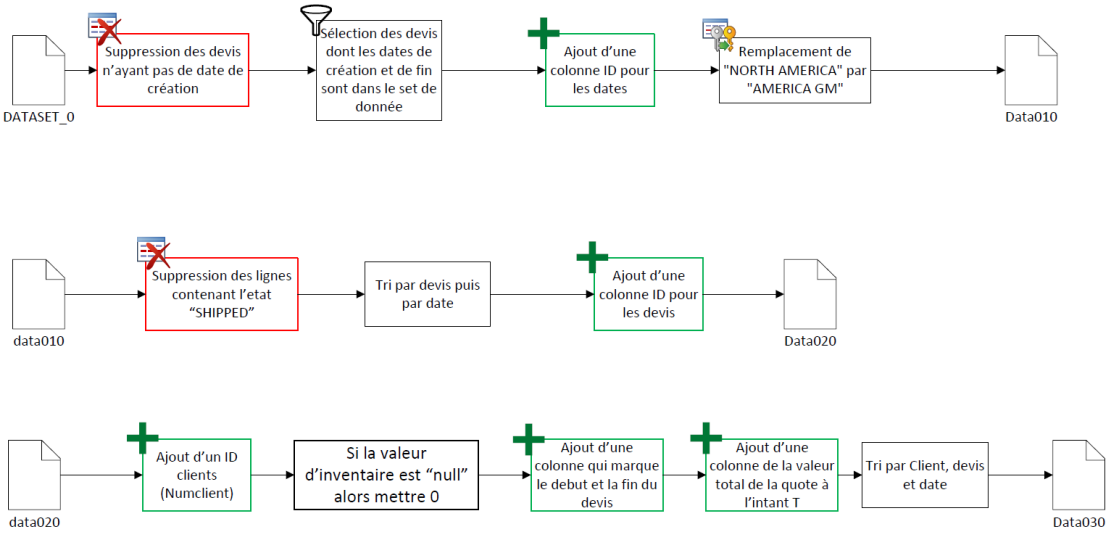

Figure 5.5 : Processus des différents nettoyages initiaux

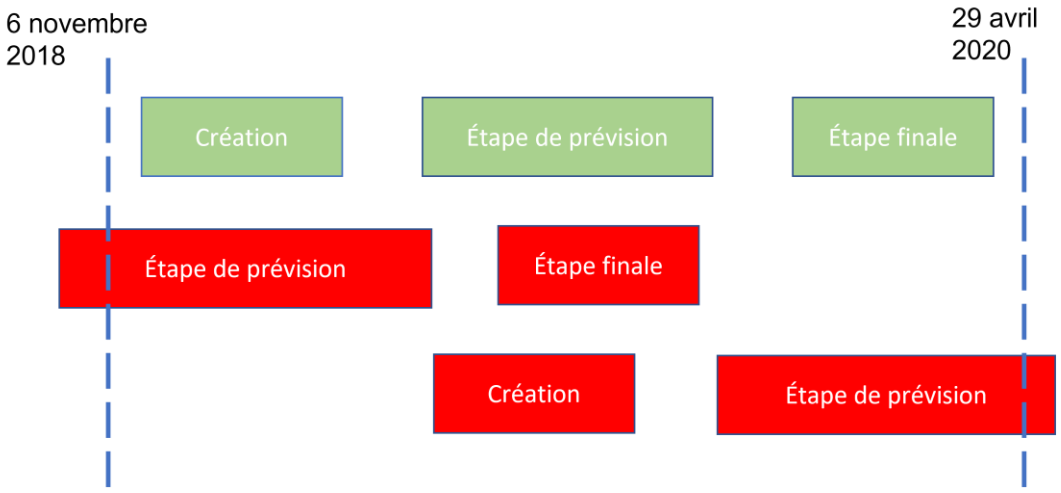

Figure 5.6 : Sélection des devis dans la bonne zone

5.1.4 Formatage des données

Pour faciliter l'étude, une colonne « IDdate » a été créée. Celle-ci va de 1 (6 novembre 2018) à 75 (29 avril 2020).

Dans la colonne « Région de planification », nous avons harmonisé les noms des régions. Par exemple, la région "AMERICA GM" étant le nouveau nom de la région "NORTH AMERICA", nous avons harmonisé les noms avec « AMERICA GM » correspondant ensuite à la région 1.

Ensuite, les devis étant vendus à partir de l'état « ORDRE DE VENTE », nous avons choisi de ne pas considérer leurs états suivants. Notre étude se concentre donc seulement sur les différents états entre la création de devis et l'atteinte de son état final (annulé ou vendu) avant leur expédition. En effet, les états du cycle de vie après « ORDRE DE VENTE » n'apportent pas plus d'information utile à la prédiction.

De plus, les données ont été triées par devis et date afin de travailler chaque devis indépendamment et de faciliter le suivi de son parcours. Aussi, afin de faciliter les analyses et éviter la gestion des noms complexes des devis, nous les avons renommés entre 1 et 41 946 dans la colonne « ID-Devis ».

La dernière étape de ce nettoyage des données a consisté à créer une colonne « Numclient ». Cette dernière correspond à un ID définissant le niveau le plus bas de la hiérarchie des clients de l'entreprise. Cette colonne est utilisée pour simplifier la valeur de la colonne « DP Sub Cust Top Client », et varie entre 1 et 136. Ces 136 clients correspondront à l'échelle la plus petite lors de l'étude des devis. En dessous, le nombre de devis et de produits commandés est trop faible pour avoir un intérêt industriel.

5.1.4.1 Marquage de séparation entre les devis et entre les dates

Pour faciliter la lecture des données, deux marquages ont été ajoutés. La colonne « débutfin » indique le début et la fin d'une « photo » d'un devis. La valeur 0 est utilisée pour marquer le début (création) d'un devis, alors que la valeur 1 est utilisée pour marquer son état final. Les « photos » dont cette colonne indique 3 est une « photo » intermédiaire du devis. Le Tableau 5.1 présente un exemple de ce marquage.

Ce marquage du début et de fin de la « photo » de chaque devis permet simplifier le calcul de sa valeur en dollars. En effet il suffit d'additionner la valeur de produit entre un 0 et un 1. Cette information est stockée dans la colonne appelée « Valeur totale ». Elle permet à chaque « photo » de savoir la valeur totale du devis.

Tableau 5.1 : Marquage début-fin

Date photo	Devis	Gamme de produit	…	DDCC	Numéro Client	débutfin
2018-11-06	ASIENDH7	1	…	Siebel 51 % to 79 %	6789	**0**
2018-11-06	ASIENDH7	2		Siebel 51 % to 79 %	6789	**3**
2018-11-06	ASIENDH7	3	…	Siebel 51 % to 79 %	6789	**1**
2019-03-15	ASIENDH7	1	….	PRÉVISION	6789	**0**
2019-03-15	ASIENDH7	2	….	PRÉVISION	6789	**3**
2019-03-15	ASIENDH7	3	….	PRÉVISION	6789	**1**
2019-06-15	ASIENDH7	1	…	ORDRE DE VENTE	6789	**0**
2019-06-15	ASIENDH7	2	…	ORDRE DE VENTE	6789	**3**
2019-06-15	ASIENDH7	3	…	ORDRE DE VENTE	6789	**1**

À la suite des différentes corrections et sélections des données, nous avons finalement retenu 44 % des devis (environ 41 000 devis), soit 32 % des données initiales (6.2 millions de lignes au lieu de 19.7 millions). L'impact des différentes préparations est visible dans la Figure 5.7. Le nom des étapes correspond au nom de la base créée visible dans la Figure 5.5.

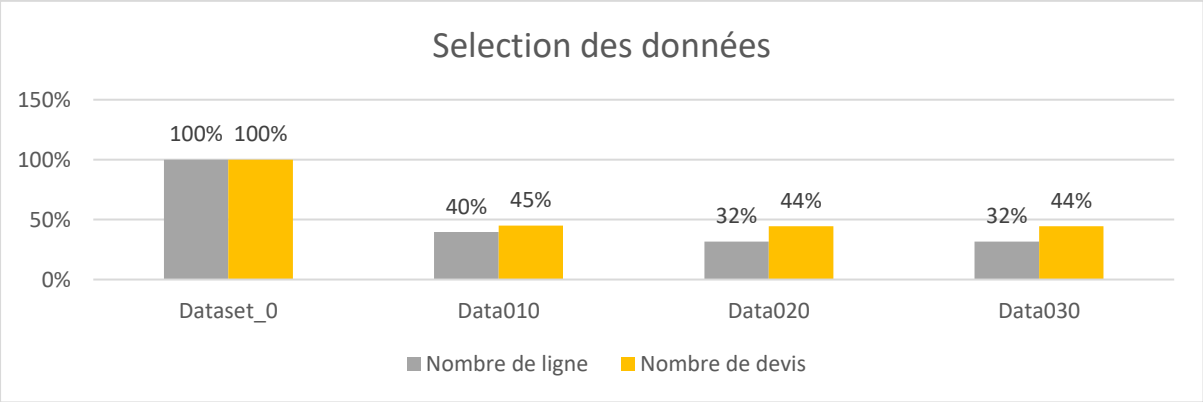

Figure 5.7 : Pourcentage de données restantes après sélection

5.1.5 Création des bases de données de travail

Pour créer la base de données de travail, nous avons utilisé les deux bases de données nettoyées, soit la base de données des devis sans doublons ni répétitions, et celle contenant les informations sur les produits. La Figure 5.8 présente le processus en question.

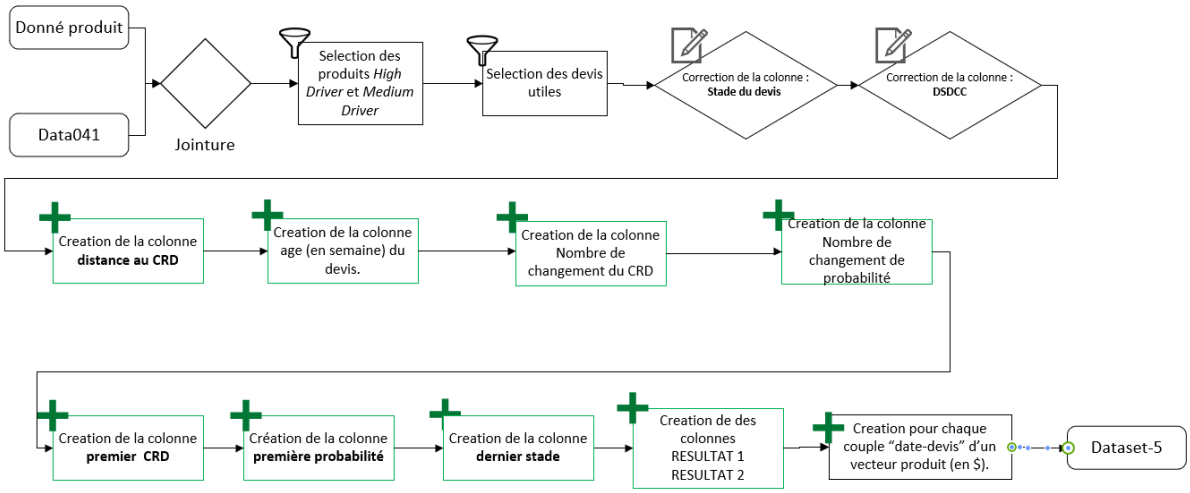

Figure 5.8 : Processus de création d'une base de données de travail

Dans un premier temps, nous avons établi un lien en elles afin d'ajouter les informations supplémentaires sur les produits, et particulièrement les informations sur l'importance stratégique des produits à notre base de données des devis.

Pour cela, on commence par appliquer deux filtres à notre nouvelle base de travail selon le processus suivant :

- Sélectionner les lignes « produits » ayant un indicateur « high driver » ou « medium driver ». Cela permet de simplifier la base et d'exclure les produits qui ne sont pas stratégiques pour l'entreprise. Nous obtenons alors 48 regroupements de produits à analyser au lieu de 63 et de supprimer 1 000 000 de lignes.
- Sélectionner les devis/dates utiles. Ce filtre exclut les devis qui ne sont pas utiles dans la création de modèles prédictifs. Ce filtre fonctionne de la façon suivante :
    - Suppression des devis en ventes directes, car inutiles pour la classification.
    - Supprimer les dates inutiles qui correspondent à une lenteur de suppression dans base. En effet, parfois des devis restent dans leur état final pendant plusieurs semaines sans disparaître. On exclut donc ces lignes.

Notre base de données évolue alors du modèle du Tableau 5.2 au modèle du Tableau 5.3. Les lignes en couleur sont éliminées.

Tableau 5.2 : Données avant modifications

Date photo	Devis	Gamme de produit	…	DDCC	Portfolio
2018-11-06	ASIENDH7	1	…	Siebel 51 % to 79 %	Low
2018-11-06	ASIENDH7	2		Siebel 51 % to 79 %	medium
2018-11-06	ASIENDH7	3	….	Siebel 51 % to 79 %	high
2019-03-15	ASIENDH7	1	….	FORECASTED	Low
2019-03-15	ASIENDH7	2	….	FORECASTED	medium
2019-03-15	ASIENDH7	3	….	FORECASTED	high
2019-06-12	ASIENDH7	1	….	SALES ORDERS	Low
2019-06-12	ASIENDH7	2	….	SALES ORDERS	medium
2019-06-12	ASIENDH7	3	….	SALES ORDERS	high
2019-06-19	ASIENDH7	1	….	SALES ORDERS	Low
2019-06-19	ASIENDH7	2	….	SALES ORDERS	medium
2019-06-19	ASIENDH7	3	…	SALES ORDERS	high

Tableau 5.3 : Données après modifications

Date photo	**Devis**	**Gamme de produit**	**…**	**DDCC**	**Portfolio**
2018-11-06	ASIENDH7	2		Siebel 51 % to 79 %	medium
2018-11-06	ASIENDH7	3	….	Siebel 51 % to 79 %	high
2019-03-15	ASIENDH7	2	….	FORECASTED	medium
2019-03-15	ASIENDH7	3	….	FORECASTED	high
2019-06-12	ASIENDH7	2	….	SALES ORDERS	medium
2019-06-12	ASIENDH7	3	…	SALES ORDERS	high

L’étape suivante consiste à appliquer une correction à deux colonnes selon le processus suivant :

- Sur la colonne « Données DC consolidé » (DDCC):
    - Remplace les valeurs « PREVISION » par « ORDRE DE VENTE » quand le stade est égal à « FERME GAGNE ». Ce changement correspond à la solution évoquée au-dessus (cf Section 5.1.2.2).
- Sur la colonne « Stade du devis » :
    - Remplacer les valeurs par leur correspondance du Tableau 4.1.
        - Par exemple :
            - « FERME PERDUE » est remplacé par « ANNULE »
            - « » est remplacé par « VENDU »

- « SOLUTION EN DÉVELOPPEMENT » est remplacé par « NEGOCIATION »

On ajoute ensuite différentes colonnes potentiellement utiles au calcul des prédictions :

- La colonne « Age » permet de connaître l'âge en semaine du devis à tout moment et donc d'étudier s'il a un impact sur son état final. A chaque « photo », l'âge du devis est incrémenté de 1.
- La colonne « Nombre de changement CRD » nous renseigne à l'instant *t* sur le nombre de fois où le devis a changé de date de fin prévue.
- La colonne « Nombre de changements de probabilité » nous renseigne à l'instant *t* sur le nombre de fois où le devis a changé de probabilité.
- La colonne « Première probabilité » renseigne sur la probabilité à la création.
- La colonne « distance au CRD » est calculée en nombre de mois avant la date de livraison prévue à ce moment *t*. La colonne permet de connaître le nombre de mois qu'il reste au devis avant d'être théoriquement vendu.
- La colonne « RÉSULTAT 1 » est binaire. La valeur est 1 si le devis est vendu, et 0 sinon. Il s'agira de la classification dite « simple ».
- La colonne « RÉSULTAT 2 » peut prendre 6 valeurs différentes (Tableau 5.4) Il s'agira de la classification dite « détaillé ».

Tableau 5.4 : Valeurs possibles par la colonne RÉSULTAT 2

Valeur	Signification
Q-1	Vendu 1 quart avant celui prévu initialement
Q0	Vendu dans le quart prévu initialement
Q1	Vendu 1 quart après celui prévu initialement
Q2	Vendu 2 quarts après celui prévu initialement
Q3	Vendu 3 quarts (ou plus) après celui prévu initialement.
Q4	Le devis est annulé

Ensuite, nous avons ajusté la structure de la base de données pour les produits (du Tableau 5.5 au Tableau 5.6), en supprimant les deux colonnes « Valeur d'inventaire en $ » et « KFU CODE » pour les remplacer par 48 colonnes. Chacune de ces 48 colonnes représente un des produits listés

initialement dans la colonne « KFU CODE ». Les valeurs mises dans ces colonnes sont celles de la colonne « Valeur d'inventaire en $ » du produit en question (cf. 4.1.1.3).

Tableau 5.5 : Structure de la base de données avant ajustement.

ID date	Devis	KFU CODE	…	DDCC	Valeur d'inventaire ($)
1	ASIENDH7	1		Siebel 51 % to 79 %	500
1	ASIENDH7	2	…	Siebel 51 % to 79 %	200
2	ASIENDH7	1	….	FORECASTED	500
2	ASIENDH7	2	….	FORECASTED	200
3	ASIENDH7	1	….	SALES ORDERS	500
3	ASIENDH7	2	….	SALES ORDERS	200

Tableau 5.6 : Structure de la base de données après ajustement.

ID date	Devis	KFU CODE 1	KFU CODE 2	….	KFU CODE 48	….	DDCC
1	ASIENDH7	500$	200$		0$		Siebel 51 % to 79 %
2	ASIENDH7	500$	200$	….	0$	….	FORECASTED
3	ASIENDH7	500$	200$	…	0$	…	SALES ORDERS

Le couple ID-Devis devient alors un identifiant unique que l'on trouve dans la colonne « IDquote ».

Finalement, la base de données finale est aussi filtrée pour exclure les colonnes inutiles pour la suite du projet. On obtient alors une base avec 88 513 lignes et 74 colonnes pour 11 927 devis différents et 123 clients.

5.1.6 Sélection et regroupement client

La base de données étant de grande taille, nous avons décidé de réaliser notre étude sur un ensemble plus restreint de clients au comportement d'achat différents. Nous avons donc sélectionné certains clients pertinents qui ont à la fois suffisamment de devis dans la base de données, et qui représentent aussi un volume d'achat significatif pour l'entreprise.

5.1.6.1 Sélection des clients

Étant donné la diversité de taille et du nombre de commandes des clients, nous avons appliqué le principe de Paréto (Figure 5.9) et choisi les clients qui correspondent à environ 80 % de l'ensemble des ventes. Par souci de confidentialité, les résultats sont présentés en pourcentage.

Cette sélection représente ainsi les 25 clients les plus importants du point de vue du montant total dépensé dans l'entreprise durant la période analysée.

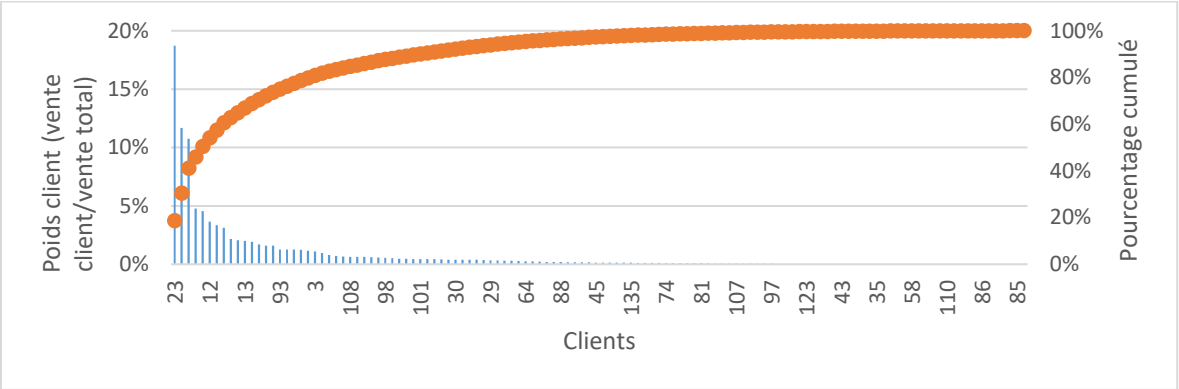

Figure 5.9 : Pourcentage de devis par client

5.1.6.2 Classification client

Étant donné que pour la période étudiée, certains des 25 clients ont peu de devis utilisables pour nos analyses (Figure 5.10), nous posons l'hypothèse que la mise en commun des devis de clients ayant des comportements d'achat similaires pourrait améliorer la qualité des prédictions de certains modèles. Par conséquent, des regroupements de clients ont été effectués de différentes façons, et à différents niveaux. Autrement dit, il s'agit de savoir si on peut obtenir une meilleure prédiction de l'état final des devis en regroupant certains clients similaires en taille ou en types de produit acheté.

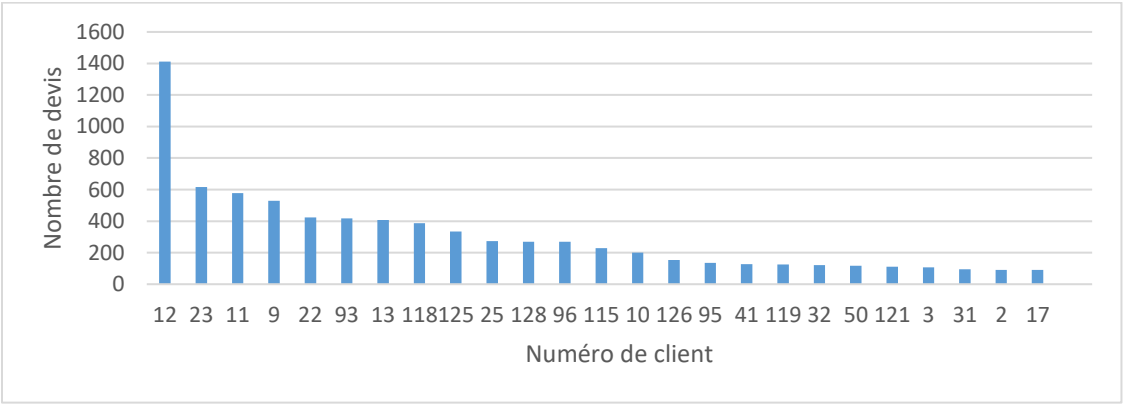

Figure 5.10 : Nombre de devis par client

Trois méthodes de regroupements ont donc été testées dans cette étude :

1. Classification par région utilisée par l'entreprise ;

2. Classification selon le volume d'achat (en dollars) du client dans l'entreprise ;
3. Classification selon le panier des produits du client ;

Les différents regroupements sont disponibles en annexe B.

5.1.6.2.1 Classification des clients par région

La première classification client reprend la hiérarchie client utilisée dans l'entreprise (Figure 5.11). Seuls 3 niveaux hiérarchiques sont possibles :

- Le client seul
- Le regroupement DP Sub Cust Top Client
- La région de planification

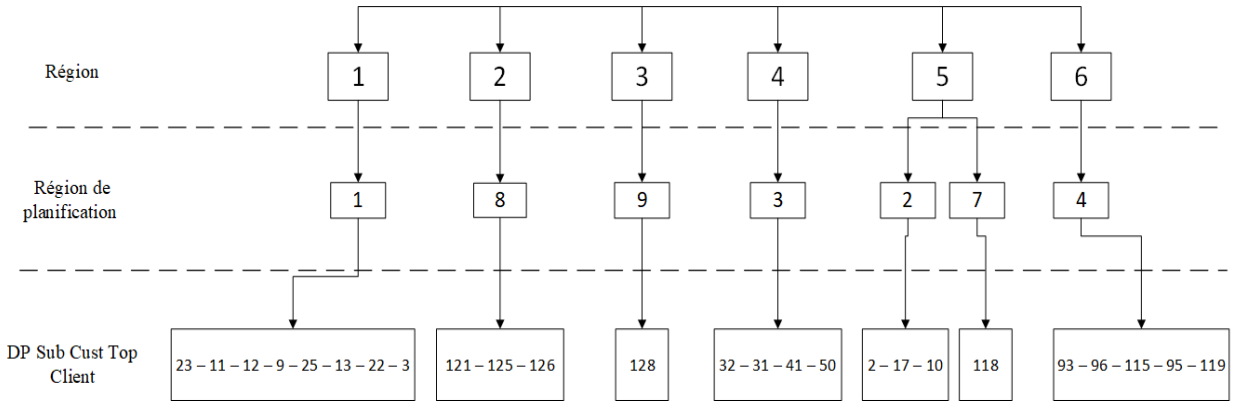

Figure 5.11 : Classifications par région

5.1.6.2.2 Classification par volume d'achat (en dollars) du client

Cette deuxième classification se base sur les volumes de vente au client au cours de la période analysée. Elle a pour but de regrouper les clients ayant un même « poids » financier pour l'entreprise. En effet, même sur l'échantillon des 25 plus gros clients, de grands écarts dans les habitudes d'achat sont visibles.

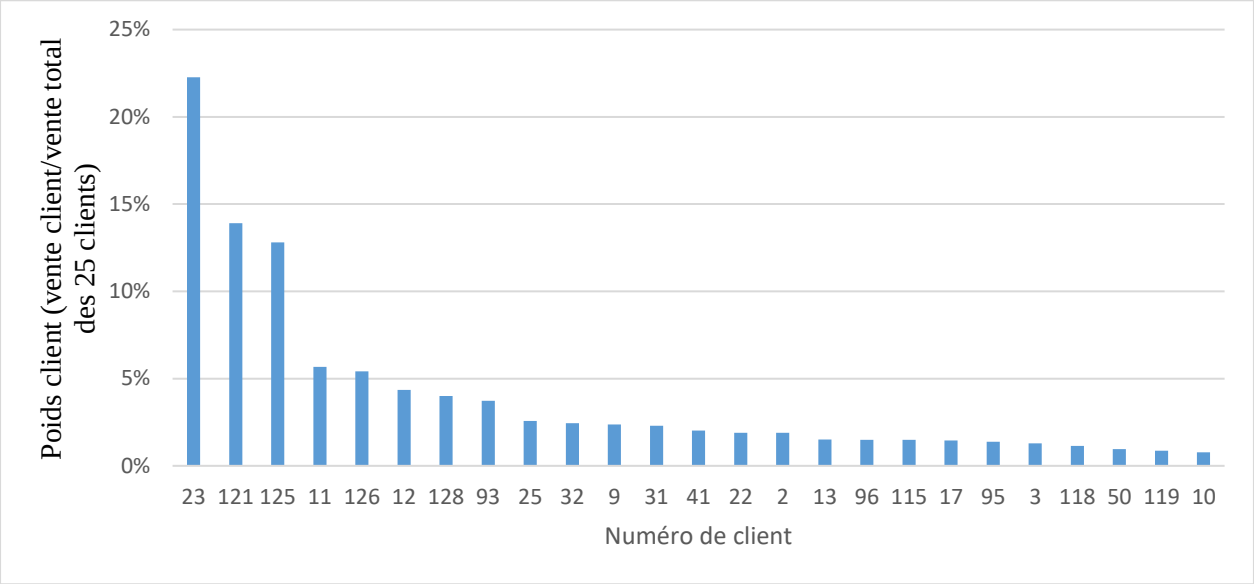

Figure 5.12 : Volume d'achat des 25 clients

Pour regrouper les clients, nous avons utilisé un critère représentant la distance euclidienne entre deux clients calculés de la façon suivante :

$$D_{argent}(C1, C2) = \sqrt{(Dep1 - Dep2)^2} \quad (4)$$

Avec $Dep1$ et $Dep2$, les volumes de ventes des clients 1 et 2 sur les 6 quarts en valeur d'inventaire.

Dans le cadre d'une analyse de regroupement réalisée à l'aide de l'algorithme de la variance minimum de Ward (ward.D2). On obtient alors un dendrogramme (Figure 5.13).

Suite à la construction du dendrogramme, nous avons utilisé le concept d'inertie [34] pour identifier des regroupements de clients intéressants (Figure 5.14). Ainsi, un grand saut d'inertie marque un changement de distance important entre les clients du regroupement. À la Figure , les trois sauts importants ont été marqués par les points vert, rouge et bleu. Ces regroupements sont identifiés par les couleurs correspondantes à la Figure 5.13. Le bleu correspond au regroupement le plus sélectif (le moins de clients par regroupement) et le vert le regroupement le plus large (le plus de clients par regroupement).

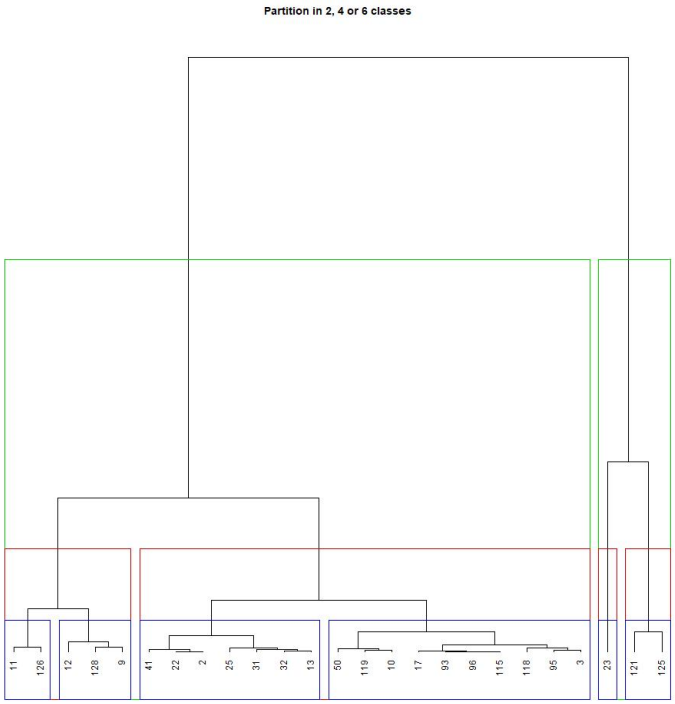

Figure 5.13 : Trois niveaux de clustering pour la distance en dollars

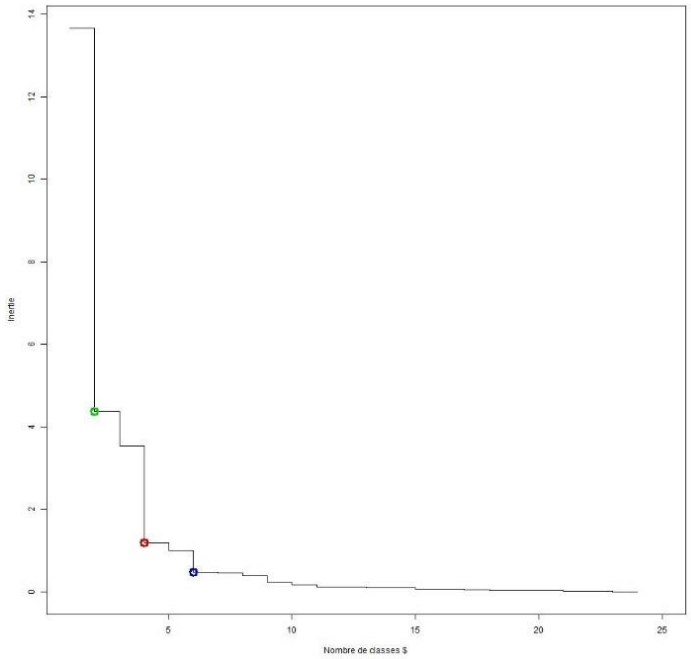

Figure 5.14 : Évolution de l'inertie selon la coupe pour la distance en dollars

5.1.6.2.3 Classification par panier des produits du client

Cette troisième classification se base sur les paniers des produits vendus aux clients au cours de la période analysée. Plus spécifiquement, nous utilisons ici la ***proportion en dollars de chaque produit vendu*** pour comparer les clients.

Cette approche de classification a pour but de regrouper les clients qui achètent les mêmes types de produits. Pour cela, on identifie chaque client par un vecteur composé des produits correspondant aux plus importants dans la classification hiérarchique de l'entreprise (soit 49 Sub KFU code). Dans les approches classifications présentées ci-dessous, nous utilisons deux méthodes pour le calcul des distances, soit la Distance euclidienne et la Distance de Manhattan, afin de voir si ces méthodes mènent à des différences significatives.

5.1.6.2.3.1 Distance produit euclidienne :

$$DproduitE(C1, C2) = \sqrt{\sum_{i=1}^{49}(P_{C1,i} - P_{C2,i})^2} \quad (5)$$

Avec $P_{C1,i}$ et $P_{C2,i}$ la proportion de vente (en % de dollars) du produit *i* chez le client C1 et C2, respectivement.

$$P_{C1,i} = \frac{Ventes_{C1,i}}{Ventes_{C1}} \quad (6)$$

$Ventes_{C1,i}$ Valeur des ventes du produit *i* au client C1 vendu sur la période étudiée (en valeur d'inventaire).

$Ventes_{C1}$, Valeur totale des ventes au client C1 vendu sur la période étudiée (en valeur d'inventaire).

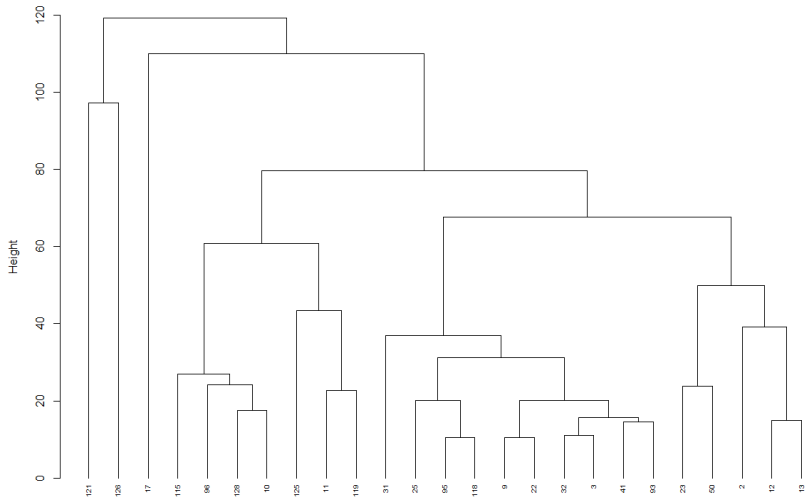

Figure 5.15 : Dendrogramme clustering client par panier de produits (distance euclidienne)

5.1.6.2.3.2 Distance produit Manhattan :

$$DproduitM(C1, C2) = \sum_{i=1}^{49} \left| P_{C1,i} - P_{C2,i} \right| \quad (7)$$

Avec $P_{C1,i}$ et $P_{C2,i}$, définis ci-dessus.

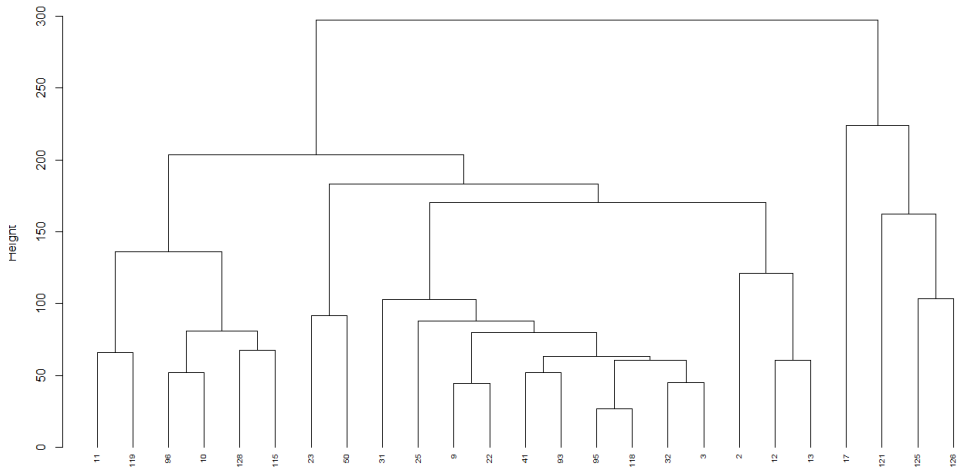

Figure 5.16 : Dendrogramme clustering client par mix produit (distance Manhattan)

Les différents dendrogrammes construits ont été soumis à l'entreprise pour les valider. Aucune anomalie n'a été détectée. Les regroupements ainsi générés ont été utilisés dans les analyses subséquentes afin d'obtenir de meilleures qualités de prédictions.

5.1.7 Présentation détaillée des 5 clients étudiés

Comme discuté plus haut, notre étude se concentre sur le développement d'outils de prédictions de l'état final des devis de 5 clients choisis pour leur diversité et leur importance stratégique pour le partenaire. Les sections suivantes présentent en détail les profils de ces clients.

5.1.7.1 Nombre de devis

Premièrement, les 5 clients sélectionnés l'ont été, car ils présentent un nombre de devis suffisants pour réaliser des analyses. Le Tableau 5.7 présente le nombre de devis que les clients ont soumis sur les 18 mois de la période étudiée.

Tableau 5.7 : Nombre de devis utilisables par client

	Nombre de devis utilisable
Client 11	577
Client 12	1411
Client 22	423
Client 23	616
Client 93	530

5.1.7.2 Proportion de devis par rapport aux ventes directes

Comme il a été discuté précédemment, les clients peuvent avoir des comportements d'achat très différents les uns des autres. Un des aspects qui différencie les clients est ainsi le volume de leur commande qu'ils passent en négociant via un devis, par rapport au volume de commandes qu'ils passent directement via un achat direct.

La Figure 5.17 montre ainsi les proportions de ventes directes (en orange) par rapport aux ventes par devis (en bleu). On peut noter premièrement que les clients sont tous différents, bien qu'ils achètent principalement par devis. Le client 23, cependant, utilise principalement les achats en ventes directes, bien qu'il reste un acheteur significatif en achat par devis. C'est pourquoi il fait tout de même partie de notre étude de cas.

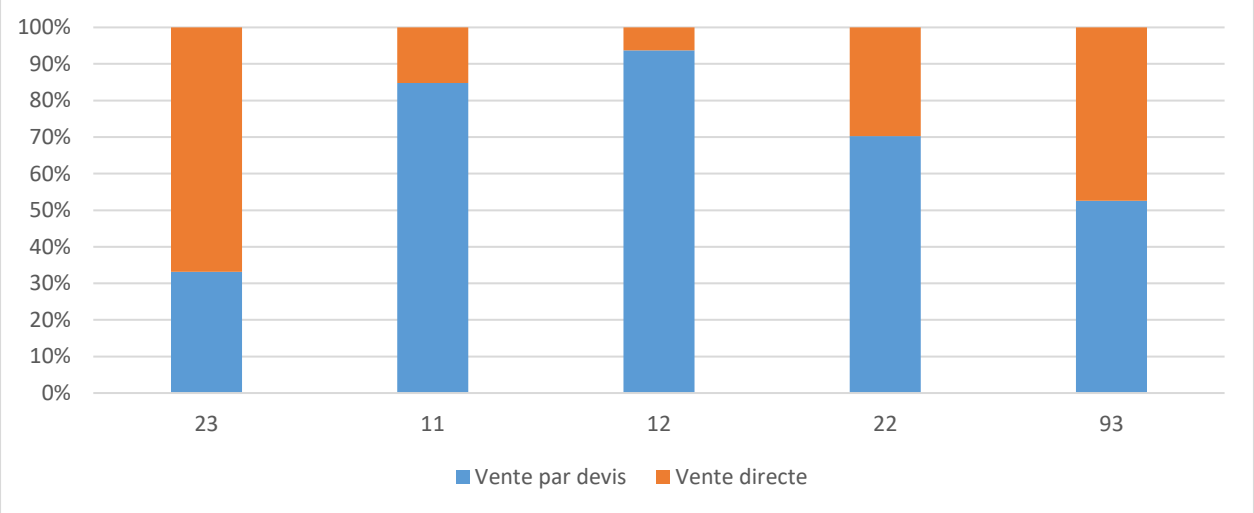

Figure 5.17 : Proportions (en dollars) ventes directes / ventes par devis pour les 5 clients

5.1.7.3 Estimation initiale de réalisation des devis

L'estimation de la réalisation des devis, aussi appelée dans ce mémoire la « probabilité du devis », est une évaluation qualitative du vendeur du potentiel de vente finale du devis. La

première estimation faite à la création du devis nous renseigne ainsi sur la première impression faite par le client. Ces dernières sont ainsi réparties de manière hétérogène entre 0 % et 100 %. Les devis des clients 11 et 23 présentent un nombre significatif de devis ayant une probabilité initiale de 80 %. Cela représente plus de 80 % des devis pour le client 11 et environ 70 % pour le client 23.

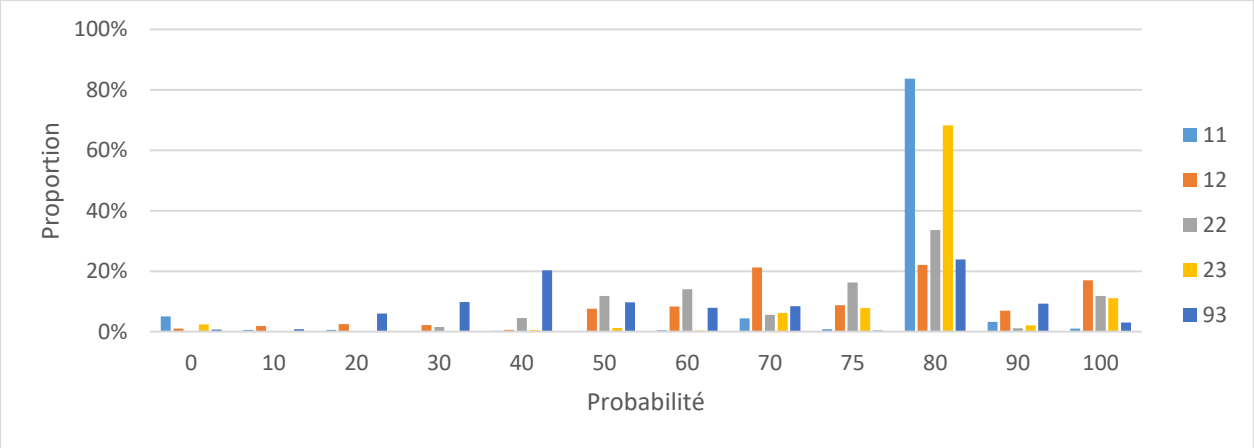

Figure 5.18 : Distribution de la fréquence de la probabilité initiale

Pour les clients 12, 22 et 93, les distributions sont cependant plus étalées. Les clients 12 et 22 présentent des probabilités initiales centrées autour de 80%. Le client 93 présente une distribution bimodale centrée autour de 80% et de 40%. (Figure 5.18)

L'analyse de ces probabilités ne permet pas de définir des comportements de client type. En revanche, elle permet de mieux comprendre le lien entre le client et le vendeur, puisque les données initiales utilisées sont issues de l'avis subjectif du vendeur.

5.1.7.4 Âge final des devis

Un autre aspect différenciant les clients est l'âge des devis lorsqu'ils arrivent à leur état final. C'est ainsi un indicateur intéressant pour caractériser le comportement client, et plus particulièrement pour savoir si un client préfère ou non conclure rapidement une vente. Il permet ainsi de connaître le temps que met un devis pour être vendu ou annulé. Par exemple (Figure 5.19), 50 % des devis du client 11 sont vendus en moins d'un mois, et 80 % en moins de deux mois. Il est donc utile d'analyser les devis au plus tôt, car la fenêtre de visibilité des cycles de vie de la plupart de ses devis est restreinte à 2 mois. Les devis qui arrivent à un âge de 44 semaines correspondent à des devis annulés, mais qui ont évolué tardivement dans cet état.

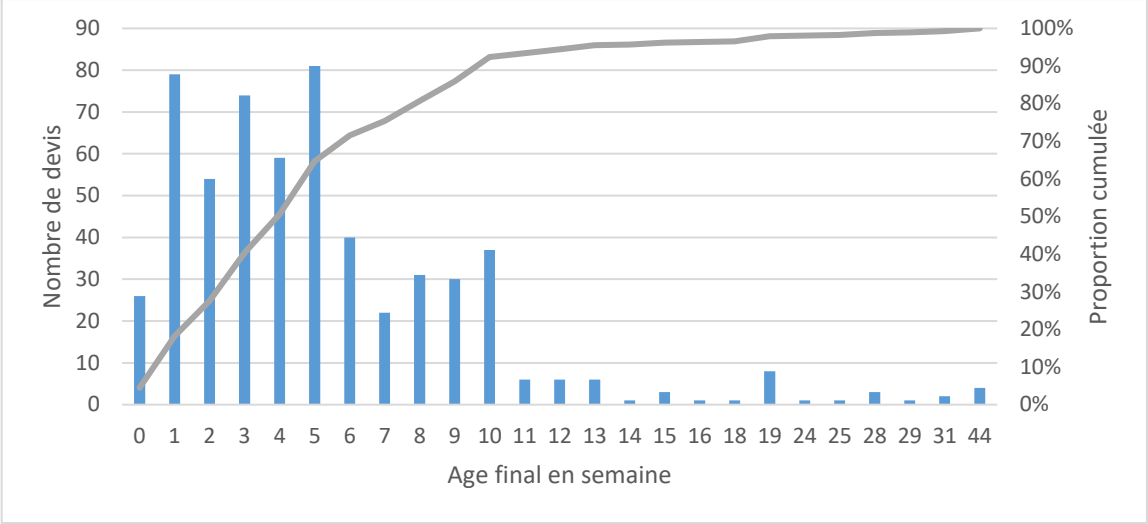

Figure 5.19 : Âge final des devis pour le client 11

5.1.7.5 Gammes de produits majoritaires

Un autre aspect caractérisant les clients concerne les lignes de produits qu'ils achètent majoritairement. Le partenaire vend ainsi plusieurs familles de technologies regroupant des sous-ensembles plus ou moins disjoints de produits compatibles entre eux. Un client peut par exemple déjà avoir investi dans une famille particulière de technologies qu'il souhaite étendre. Connaître le profil d'achat des clients est donc un aspect qui peut indiquer quelques familles de produits il peut acheter dans le futur. Par exemple (Figure 5.20), les clients 11 et 23 présentent une gamme de produits relativement pointue. Le client 11 achète une majorité de produits 6, et le client 23 une majorité de produits 3. Le profil d'achat des autres clients est plus diversifié.

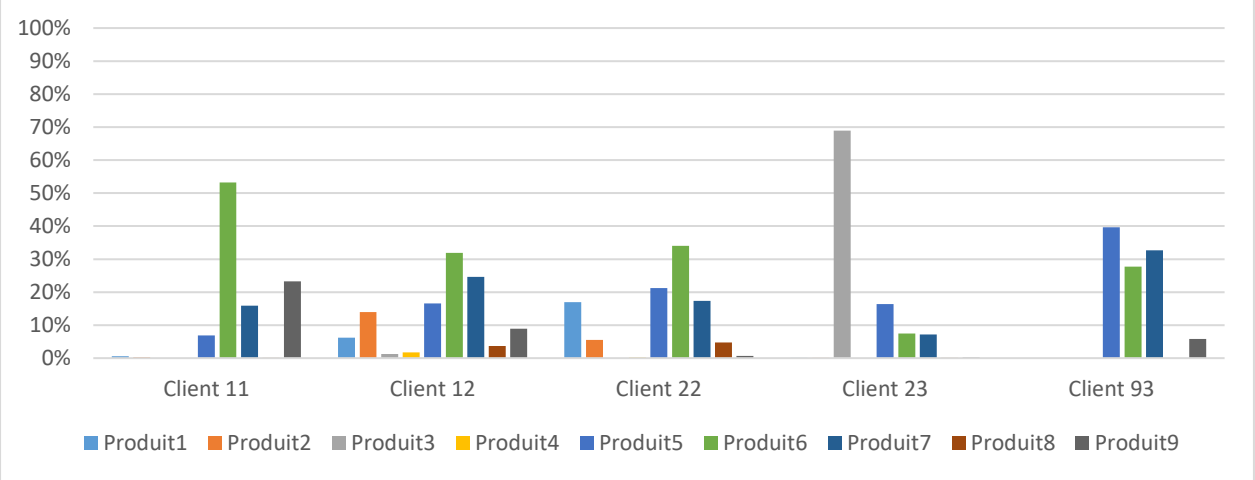

Figure 5.20 : Distribution des achats par gamme de produits

5.1.7.6 Distribution de l'état final des devis étudiés

Finalement, un dernier aspect important caractérisant le profil des clients concerne la distribution de l'état final de leurs devis. Comme nous l'avons expliqué au Tableau 5.4, l'état final d'un devis va de Q-1 (vendu 1 quart avant la date prévue), Q0 (vendu à la date prévue) jusqu'à Q3 (vendu 3 quarts après la date prévue), et Q4 (annulé).

La majorité des clients sélectionnés ont ainsi une majorité de leurs devis qui se réalise à la date (« CRD ») initialement prévue (état Q0). L'état Q4, « DEVIS ANNULE », représente aussi généralement le second état final le plus fréquent. Par exemple (Figure 5.21), le client 22 a un grand nombre de devis annulés, à la fois en nombre et en valeur. À la vue des proportions, il faudra bien regarder les résultats pour chacune des classes. Les indicateurs globaux ne seront pas toujours pertinents. Le cas des devis annulés est important, on essaiera de prévoir au mieux l'annulation.

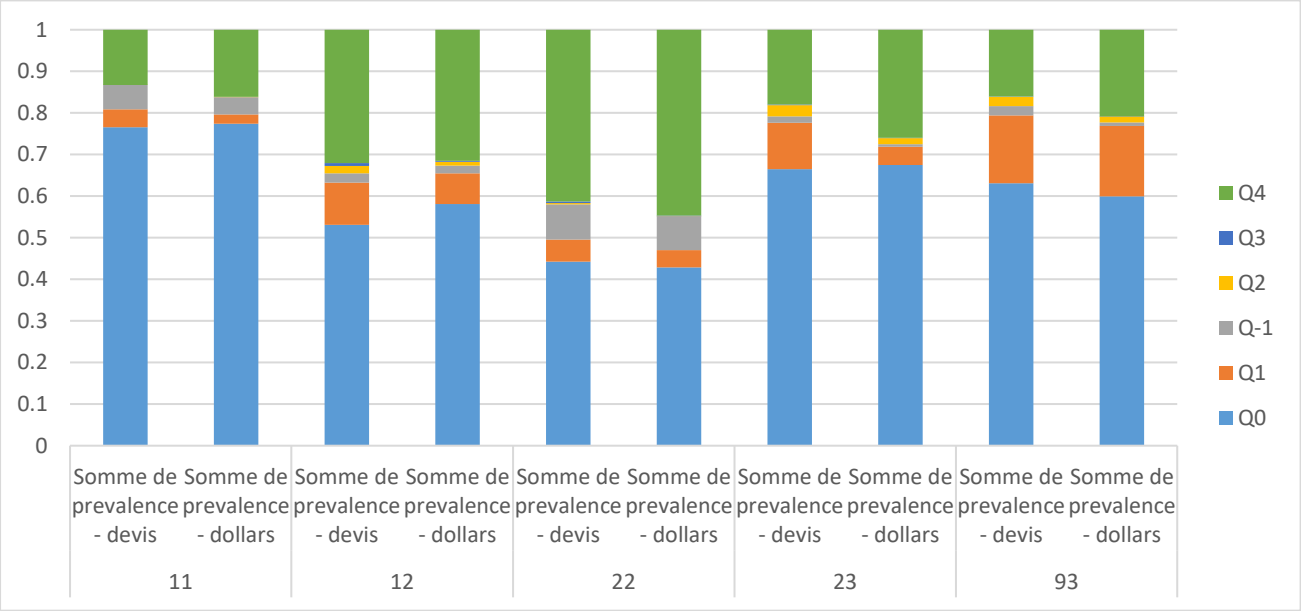

Figure 5.21 : Distribution des états finaux des devis en nombre de devis et en dollars pour les 5 clients étudiés.

5.2 Modèles de prédiction de l'état final des devis

Dans cette partie, nous présentons les derniers ajustements réalisés à la base de données, et les deux méthodes retenues de prédiction par classification des devis. Autrement dit, les méthodes de

prédiction que nous proposons d'étudier ici sont des approches permettant de prévoir l'état final discret du devis à partir de ses caractéristiques courantes. Il s'agit donc de méthodes de classification. De plus, nous avons aussi évalué deux modélisations des états finaux, soit une modélisation simple (devis vendu ou devis annulé), et une modélisation détaillée tenant compte du retard possible de la vente par rapport à la date demandée initiale par le client.

5.2.1 Méthode de validation croisée

La méthode de validation croisée utilisée dans ce projet dérive de la méthode classique de validation croisée de type *k-fold*. Étant donnée la nature spécifique des données, il a fallu adapter cette méthode. Le processus général de cette validation croisée contient les grandes étapes décrites dans les sous-sections suivantes et dans la Figure 5.22.

5.2.1.1 Construction des bases de données de travail des modèles étudiés

Les modèles de prédictions que nous avons choisi d'étudier sont dédiés à un client à la fois. Par conséquent, la base de données générale de travail a dû être divisée par client et regroupement de clients tels que décrits précédemment. Chaque modèle étudié a donc sa propre base de données de travail spécifique à la fois pour un client et un regroupement de clients.

Chaque point de données individuel contenu dans ces bases de données de travail spécifiques représente une « photo » d'un devis à un moment donné. Cette « photo » contient de l'information générale sur le devis (client, produits-quantité, date due, etc.), son état à un moment donné (âge, % de confiance estimée par le vendeur), ainsi que son état final qui est la variable dépendante que nous cherchons à prédire. Le cycle de vie de chaque devis est donc décrit par un ensemble plus ou moins grand de « photos ». Ces bases de données de travail spécifiques contiennent donc toutes les « photos » de tous les devis des clients du regroupement concerné, ce qui implique que certains points de données sont corrélés puisqu'ils représentent des « photos » du même devis à différents moments.

Ainsi, afin d'éviter d'utiliser des points de données corrélés pour l'entraînement des modèles, nous avons dû introduire une étape de sélection aléatoire des points de données présentées à la section suivante.

5.2.1.2 Sélection aléatoire des « devis-âge »

Afin de créer les bases de données d'entraînement et de test, nous avons dû préalablement sélectionner une seule « photo » pour chaque devis. Cela a été réalisé en sélectionnant de manière aléatoire un âge pour chacun des devis. L'ajout de cette étape préalable garantit que tous les devis sont représentés une seule fois dans la sous-base de données de travail spécifique. Puisque le cycle de vie de chaque devis contient en moyenne 5 « photos », cette sélection aléatoire réduit la taille de la base de données de travail spécifique de 80 %. Cette réduction n'affecte cependant pas l'utilisabilité de ces données puisqu'elles étaient initialement suffisamment nombreuses. Pour la construction de chacun des modèles de prédiction, cette étape méthodologique mène donc à une sous-base de données spécifique utilisée dans un processus de validation croisée décrit ci-dessous.

Cependant, afin de vérifier que cette étape de sélection aléatoire n'affecte pas la performance des modèles de prédiction produits, nous avons réalisé une série d'expériences dans laquelle nous avons testé et comparé plusieurs sélections aléatoires. Ces résultats sont présentés et discutés plus loin.

5.2.2 Construction des bases de données d'entraînement et de test

Le processus général de construction des bases de données d'entraînement et de test est présenté à la Figure 5.22. Il utilise donc au départ une des sous-bases de données spécifiques de travail ne contenant qu'une seule « photo » de chaque devis concerné. À partir de cette base de données, les devis sont séparés en deux. Une partie concerne seulement les données du client étudié (base de données C), et l'autre partie concerne le reste des devis (base de données S) appartenant au regroupement de clients étudiés. Si ce regroupement ne contient que le devis du client étudié, cette étape n'est pas réalisée (car la base de données S est vide), et le processus de validation croisée mis en œuvre est le processus classique.

Si le regroupement de clients en contient plusieurs, on subdivise alors aléatoirement les deux bases de données en k=10 sections de taille similaire (base de données S_j et C_j, avec $j \in [1,10]$). Ensuite, un processus itératif pour i allant de 1 à k=10 permet d'entraîner et de tester le modèle proposé 10 fois avec des bases de données d'entraînement et de test différents à chaque fois. À chaque itération, la base d'entraînement correspond à la sous-base de données spécifique de

travail (S+C) à laquelle on retire les données des sections S_i et C_i. La base de données de test correspond alors à la section C_i. En effet, puisque l'objectif est de créer un modèle de prédiction spécifique au client étudié, seules les « photos » des devis de ce client peuvent être utilisées pour réaliser les tests.

Lors de cette étape, le modèle entraîné est aussi optimisé avant d'être testé. Cette partie est discutée ci-après. Lors des tests, on calcule les indicateurs de performances mentionnés dans la méthodologie générale, et on procède à la validation et au test suivants. Les résultats sont compilés et moyennés pour fin de comparaison.

Comme expliqué ci-dessous, cette validation croisée a été réalisée plusieurs fois afin de s'assurer que la sélection aléatoire initiale des « photos » des devis ne crée pas de biais particulier.

Le but de ces étapes est d'identifier les meilleurs paramètres des modèles et le meilleur regroupement de clients pour chaque modèle de prédiction.

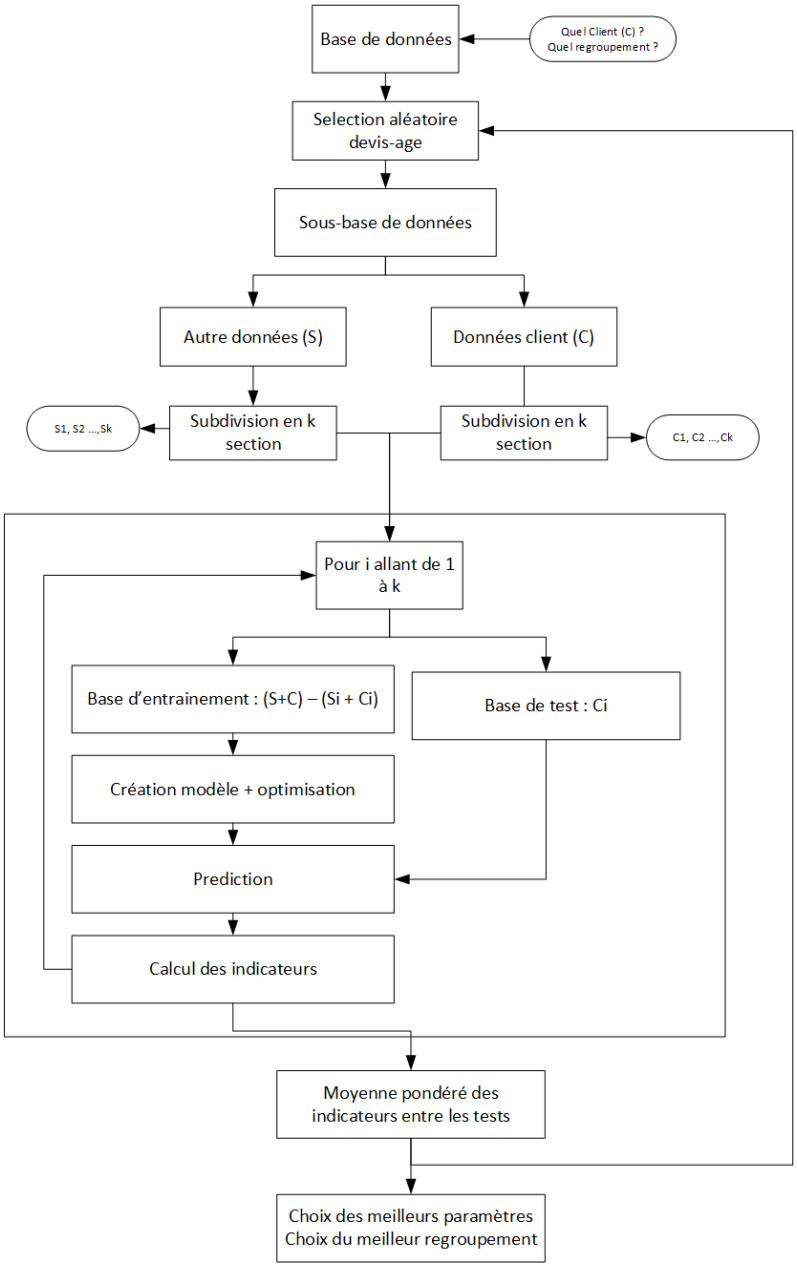

Figure 5.22 : Processus général de construction des bases de données d'entraînement et de test.

5.2.3 Modélisation de l'état final des devis (variable dépendante à prédire)

Comme nous l'avons mentionné ci-dessus, nous proposons deux approches de modélisation de l'état final discret des devis.

La première approche de modélisation (simple) ne cherche qu'à savoir si le devis a été vendu. La valeur obtenue est binaire : 0 pour « ANNULE » et 1 pour « VENDU ».

La deuxième approche de modélisation (détaillé) cherche à déterminer si le devis a été vendu ou annulé, mais aussi à calculer le délai entre la date de la vente réelle et la date de la vente demandée initialement par le client.

On dispose donc de 6 valeurs différentes (Tableau 5.4) :

- Q-1 : Le devis est vendu 1 quart avant le quart prévu ;
- Q0 : Le devis est vendu sur le quart initialement prévu ;
- Q1 : Le devis est vendu 1 quart après le quart prévu ;
- Q2 : Le devis est vendu 2 quarts après le quart prévu ;
- Q3 : Le devis est vendu 3 quarts ou plus après le quart prévu ;
- Q4 : Le devis a été annulé.

5.2.4 Modélisation de l'état courant des devis (variables indépendantes)

En ce qui concerne les variables indépendantes caractérisant l'état courant des devis, de nombreux essais ont été réalisés au cours du projet. Chaque essai nous a permis de mieux comprendre la pertinence et les limites de chaque variable. Nous avons donc ajusté le choix des variables de nombreuses fois avant de retenir les variables suivantes :

Tableau 5.8 : Variables indépendantes utilisées dans les modèles

Nom des colonnes	Nombre de colonnes	Type	Définition
Les produits	49	Flottant positif	Valeur en dollars de chaque produit dans le devis à l'instant *t*.
Somme totale	1	Flottant positif	Valeur en dollars du devis à l'instant *t*
Probabilité du devis	1	Entier (entre 0 et 100)	Probabilité du devis à l'instant *t*
Âge du devis	1	Entier positif	Âge du devis en semaines
Changement CRD	1	Entier positif	Nombre de fois que le CRD a changé dans le passé de ce devis
Distance	1	Entier positif	Nombre de mois avant d'arriver à échéance du CRD du devis.
Première probabilité	1	Entier (entre 0 et 100)	Probabilité à la création du devis
Occurrence	1	Entier positif non nul	Nombre de doublons du devis
IDprev	1	Entier	ID du premier CRD

5.2.5 Évaluation des modèles et indicateurs

Pour évaluer nos modèles, nous avons défini différents indicateurs de comparaison, dont deux principaux :

- Le pourcentage d'erreurs de classification devis (1)
    - Calcul des indicateurs en prenant le devis seul
- Le pourcentage d'erreur de placement de dollars (2).
    - Calcul de l'indicateur en prenant en compte le poids en dollars du devis.

5.2.6 Modélisation par des arbres

5.2.6.1 Approche 1 : Arbre de décision

La première approche choisie pour classer les devis est l'arbre de décision. Pour la mettre en œuvre cette approche, nous avons utilisé la bibliothèque R « rpart ». L'indice de GINI a été utilisé par défaut dans les modèles. Il sera donc employé aussi bien dans l'arbre de décision que dans la deuxième approche (ci-dessous, forêts aléatoires).

La modélisation des arbres de décisions a été la suivante. Avec les données d'entraînement (cf. 5.2.1.1), un arbre est construit avec les paramètres de construction par défaut suivant : d'une part la division minimale d'une feuille (Minsplit) est fixée à 7 éléments et d'autre part la profondeur maximale de l'arbre est égale à 30. On prendra un paramètre de complexité (cp) égal à 0 dans l'arbre, pour obtenir un arbre non élagué.

Une fois l'arbre de classification construit, on passe à l'étape d'élagage. Il permet d'obtenir un meilleur sous-arbre, plus petit, et avec une meilleure prédiction. Cette étape permet aussi d'éviter le sur entraînement. L'arbre final est obtenu grâce à une méthode de validation croisée testant différentes versions élaguées de l'arbre initial. Un bon élagage correspond à une valeur du paramètre cp permettant d'obtenir une petite erreur de validation croisée. Dans l'exemple de la Figure 5.23, ce minimum est obtenu pour une valeur de cp égale à 0.00083. Elle est utilisée comme règle d'arrêt pour construire notre nouvel arbre.

Une fois construit, il permet d'établir les prédictions avec les données de test. Le processus recommence ensuite avec les différentes bases d'entraînements et de tests (cf. 5.2.1.1).

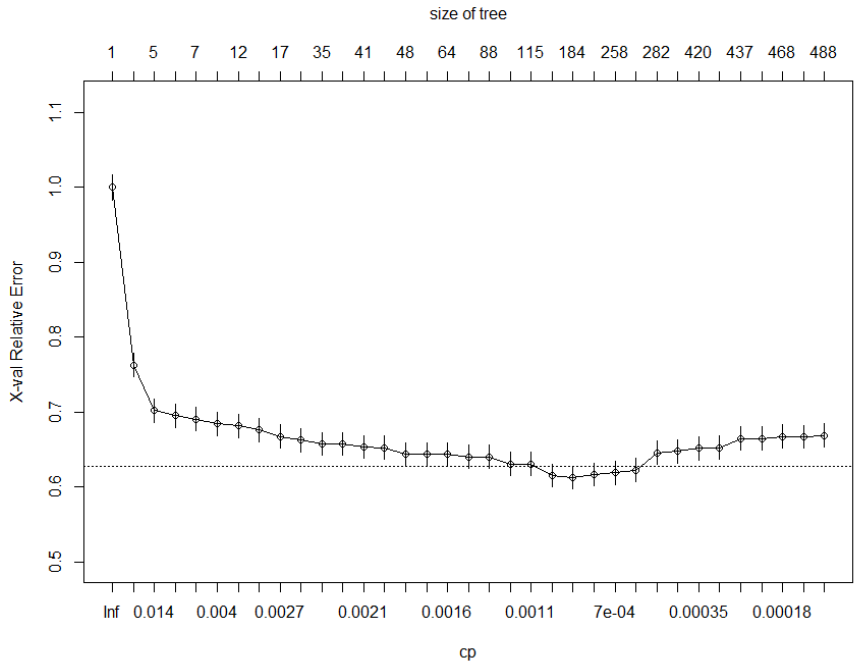

Figure 5.23 : Erreur de l'arbre en fonction de sa taille de l'arbre

5.2.6.2 Approche 2 : Forêts aléatoires

Nous allons voir dans cette section l'application de l'approche des forêts aléatoires de la bibliothèque R *RandomForest*. Cette approche a été choisie pour ses très bonnes capacités de classification. Elle est appliquée de la même manière sur les 5 clients avec les 11 agrégations possibles.

Pour la mettre en oeuvre, une forêt aléatoire est créée. Le taux d'erreur « Out Of Bag » (OOB) est une estimation de l'erreur interne d'une forêt aléatoire en cours de construction. Grâce à ce taux, le nombre d'arbres optimal est défini pour obtenir les meilleurs résultats. Le modèle est ensuite entraîné et testé avec ce nombre d'arbres. La Figure 5.24 montre un minimum d'erreur pour le modèle avec 358 arbres.

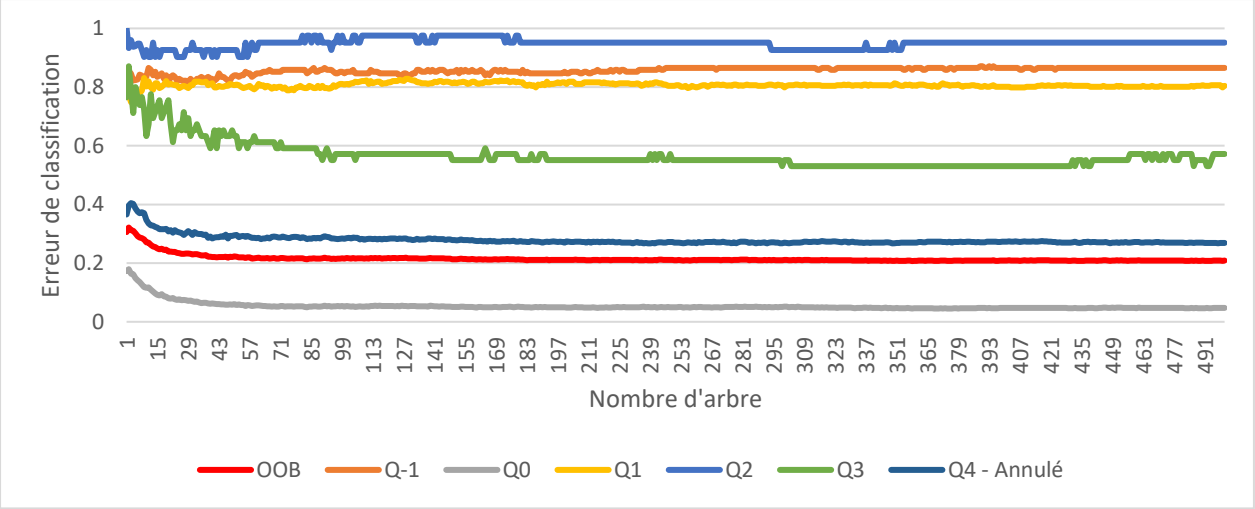

Figure 5.24 : Erreurs de classification selon le nombre d'arbres

5.2.6.3 Synthèse des modèles testés

Les analyses réalisées à l'aide des données se résument de la façon suivante :

- 2 approches d'apprentissage : arbres de décision et forêts aléatoires ;
- 2 modélisations de l'état final des devis : simple et détaillé ;
- 11 regroupements de clients ;
- 5 clients.

Nous avons donc testé 44 modèles de prédiction de l'état final des devis pour chacun des 5 clients, pour un total de 2 x 2 x 11 x 5 = 220 modèles. Chacun de ces modèles a été testé par le protocole de validation croisée décrit dans le chapitre méthodologie. De plus, afin de s'assurer que le processus de sélection aléatoire de ce protocole ne mène pas à un biais, nous avons testé 10 répétitions de sélection aléatoire pour tous les modèles de clients, ce qui correspond à 2 x 2 x 11 x 2 x 10 = 880 modèles.

Dans la suite du mémoire, **ADD** signifie Arbre de décision détaillé, **ADS** Arbre de décision simple, **FAD** forêts aléatoires détaillées et **FAS** forêts aléatoires simples.

5.3 Résultats

Dans cette section nous présentons une synthèse des résultats des différentes analyses, ainsi que les détails de l'analyse pour le client 11. Les résultats présentés sont la compilation des résultats des tests de prédiction par validation croisée. Les résultats détaillés pour les autres clients sont présentés en annexe C à G.

5.3.1 Effet de la sélection aléatoire des données

Pour évaluer le biais de sélection des données expliquée plus haut, nous avons répété les calculs de test par validation croisée avec 10 sélections aléatoires différentes. Chaque sélection aléatoire a donc été utilisée pour tester tous les modèles pour les clients 11 et 23. Pour chaque modèle, les résultats des scores F1 et de la justesse globale ont été compilés, et leurs écarts types calculés. Ces résultats sont synthétisés dans les Tableau 5.9 et Tableau 5.10. Ces tableaux montrent plus particulièrement l'écart type, d'une part, pour le score F1 de la classe annulation, et, d'autre part, pour la justesse globale selon les 11 regroupements pour les clients 11 et 23.

Ainsi, nous pouvons observer que les écarts types sont très faibles pour la très grande majorité des tests. Par conséquent, on peut affirmer que la sélection aléatoire des « devis-âge » n'impacte que légèrement les résultats. Cependant, une approche d'apprentissage (arbre de décision) pour la modélisation détaillée de l'état final des devis (ADD) est parfois impactée et présente des écarts allant de 0 % à 32 % pour le score F1. Cette anomalie est similaire pour les 2 clients testés.

Ces résultats sont intéressants à plusieurs égards. Premièrement, concernant la justesse globale, tous les modèles testés montrent que la sélection aléatoire impacte peu les résultats (faible écart type). Deuxièmement, pour la grande majorité des modèles testés, ces résultats semblent aussi montrer que ces modèles sont peu sensibles au surapprentissage puisque, quelle que soit la sélection aléatoire, la qualité des prédictions sont similaires. Troisièmement, pour le score F1, l'approche par arbres de décision avec une modélisation détaillée est plus sensible au surapprentissage puisque les résultats sont plus hétérogènes, mais similaires pour les 2 clients testés.

Tableau 5.9 : Écart type du score F1 (annulation) et justesse globale selon le regroupement (Client 11)

Regroupement	1	2	3	4	5	6	7	8	9	10	11
Écart type de score F1 (annulation)											
ADD	5%	3%	32%	4%	4%	20%	4%	4%	0%	3%	0%
ADS	3%	4%	2%	3%	3%	2%	3%	3%	3%	5%	2%
FAD	3%	2%	2%	4%	3%	3%	3%	3%	3%	3%	2%
FAS	3%	4%	3%	3%	3%	3%	4%	2%	2%	4%	3%
Écart type de la justesse globale											
ADD	2%	1%	4%	1%	1%	2%	1%	1%	1%	1%	1%
ADS	1%	1%	0%	1%	1%	0%	1%	1%	0%	2%	0%
FAD	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%
FAS	1%	1%	1%	1%	1%	1%	1%	0%	0%	1%	1%

Tableau 5.10 : Écart type du score F1 (annulation) et justesse globale selon le regroupement (Client 23)

Regroupement	1	2	3	4	5	6	7	8	9	10	11
Écart type de score F1 (annulation)											
ADD	4%	23%	23%	4%	5%	3%	4%	17%	18%	3%	3%
ADS	4%	5%	4%	3%	4%	6%	3%	5%	5%	4%	3%
FAD	4%	3%	3%	2%	2%	3%	2%	4%	4%	1%	1%
FAS	4%	3%	3%	3%	4%	3%	3%	4%	4%	2%	3%
Écart type de la justesse globale											
ADD	2%	4%	4%	2%	3%	1%	2%	3%	4%	2%	2%
ADS	1%	1%	1%	1%	1%	2%	1%	1%	2%	1%	1%
FAD	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%
FAS	1%	1%	1%	1%	1%	1%	1%	1%	1%	0%	1%

5.3.2 Synthèse des résultats

Cette section présente une synthèse des résultats des 220 modèles testés. Les résultats sont ainsi présentés par regroupement, par modèle et par indicateurs

5.3.2.1 Justesse globale

La Figure 5.25 présente les résultats de la justesse globale en dollars pour le client 11 selon le regroupement des clients (de 1 à 11) pour les approches d'apprentissage et de modélisation de l'état final choisies. La justesse est un indicateur intéressant. Il permet d'évaluer les modèles dans leur globalité. Cet indicateur donne le pourcentage de bonne classification. Cependant, il est

difficile de conclure seulement avec cet indicateur, les résultats sont trop proches. Nous devons utiliser d'autres indicateurs en complément.

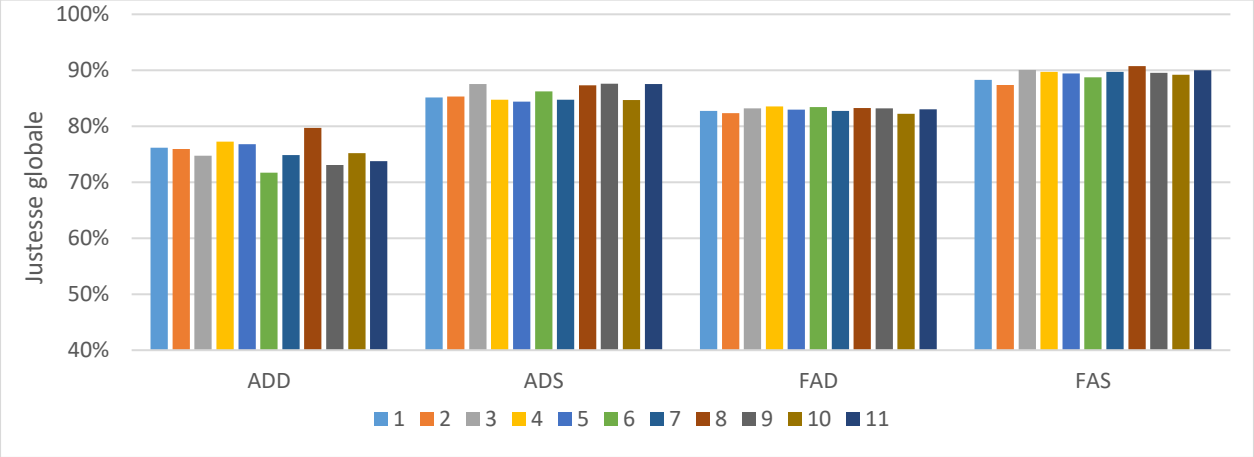

Figure 5.25 : Arbre de décision simple – Justesse globale en dollars – Client 11

5.3.2.2 Précision, sensibilité, score F1, spécificité et justesse globale

Afin de réaliser une analyse plus poussée, la qualité des modèles de prédiction a été étudiée de plusieurs façons. Premièrement, nous avons analysé les indicateurs de performance uniquement sur la classe annulation du devis. En effet, nous cherchons dans un premier temps à mesurer la qualité des modèles à prévoir l'annulation des devis. Deuxièmement, nous avons analysé les indicateurs de performance sur les autres classes. Autrement dit, nous cherchons ici à mesurer la qualité des modèles à prévoir la vente des devis ainsi que les quarts où la vente se réalise.

5.3.2.2.1 Classe annulation

La Figure 5.26 présente donc les niveaux de précision de la prédiction de l'annulation des devis pour chaque regroupement des clients selon les modèles étudiés. On peut ainsi visualiser l'impact du choix du modèle et du regroupement sur la précision de la classe annulation. La précision permet de savoir le taux de faux positif. Une précision de 100 % signifie que 100 % des devis prédits dans la classe « ANNULÉ » le sont réellement.

Pour le client 11, on constate ainsi de grands écarts de précision selon les choix faits. L'approche d'apprentissage présentant de bons résultats pour la plupart est l'approche des forêts aléatoires en utilisant une modélisation simple de l'état final des devis (« VENDU » ou « ANNULÉ »). À

l'opposé, l'approche des arbres de décision présente une hétérogénéité des résultats, bien que certains regroupements présentent d'excellents résultats.

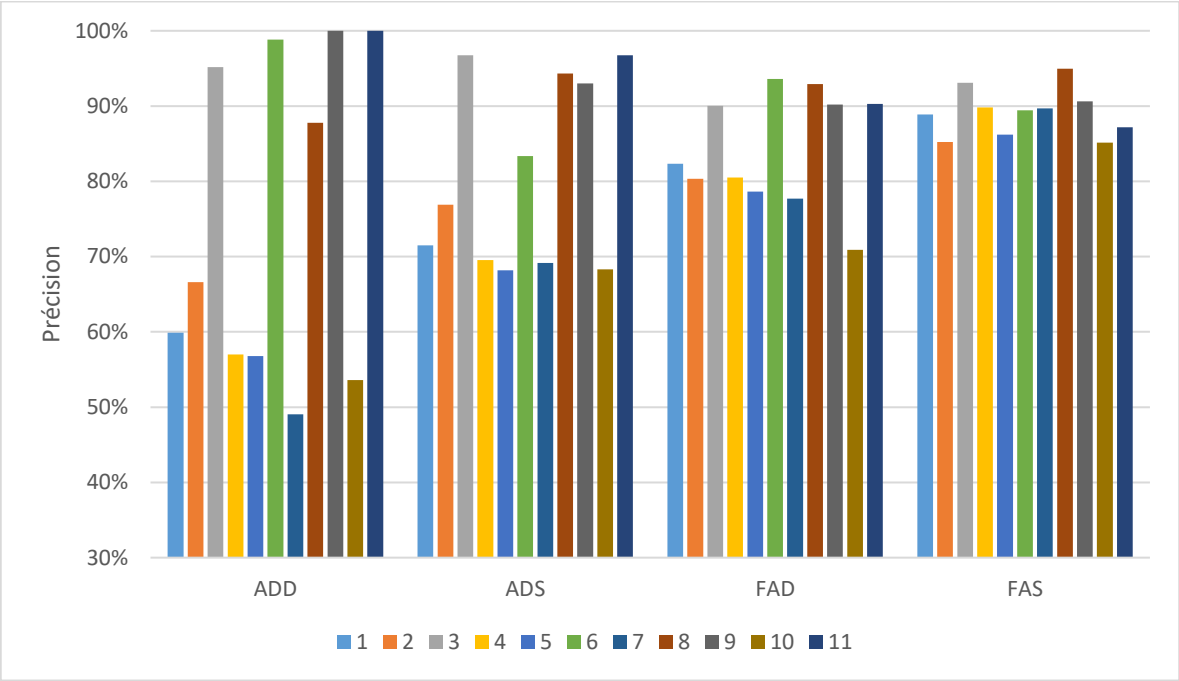

Figure 5.26 : Précision en fonction du modèle et du regroupement (1 à 11) pour la classe annulation (en dollars)

Les Figure 5.27 (a) et (b) permettent de visualiser la distribution des quatre modèles selon trois indicateurs. Le graphique de gauche (a) présente l'évolution de la précision du modèle selon sa justesse. On peut y noter que les modèles forêts aléatoires (en orange et jaune) sont plus regroupées.

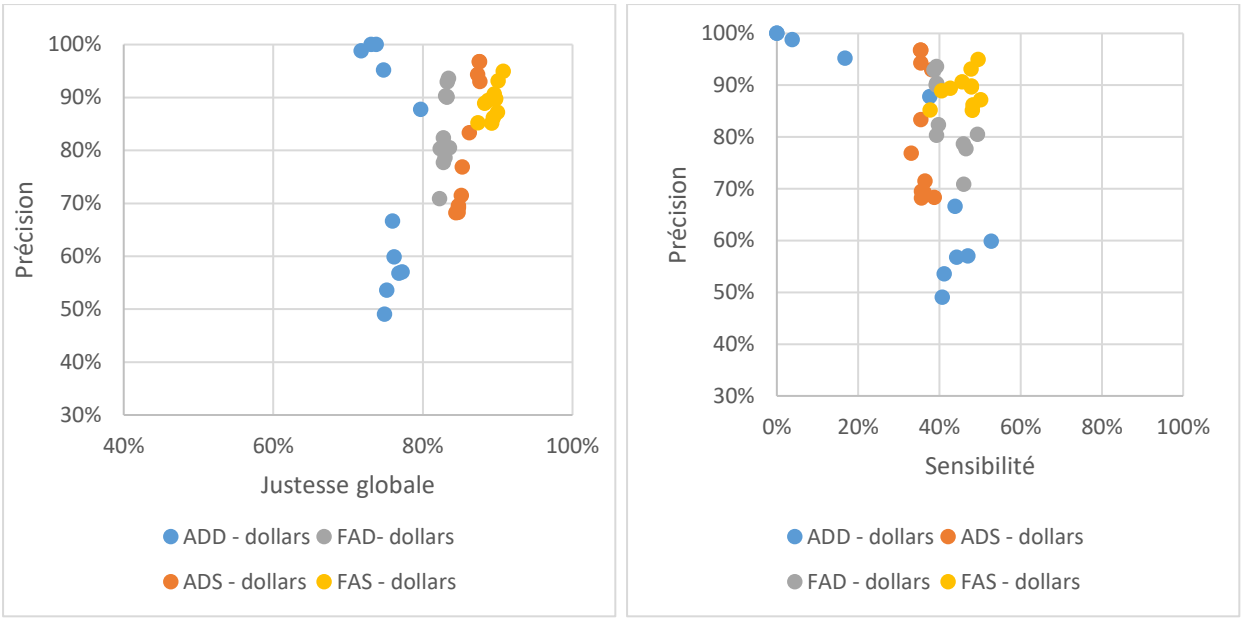

Figure 5.27 : Précision de la classe annulation en fonction de la justesse globale (a) à gauche et en fonction de la sensibilité (b) à droite pour le client 11 (en dollars)

On peut également visualiser sur le graphique de droite (b) la sensibilité selon la précision. Les meilleurs résultats sont visibles pour le modèle FAS. On peut notamment remarquer que l'on obtient relativement peu de faux positifs (précision élevée), mais un nombre significatif de faux négatifs (sensibilité faible). Cela signifie que les devis prédits comme étant annulés le sont souvent en réalité. Cependant, beaucoup de devis annulés n'arrivent pas dans la bonne classe, et sont ainsi mal prédits.

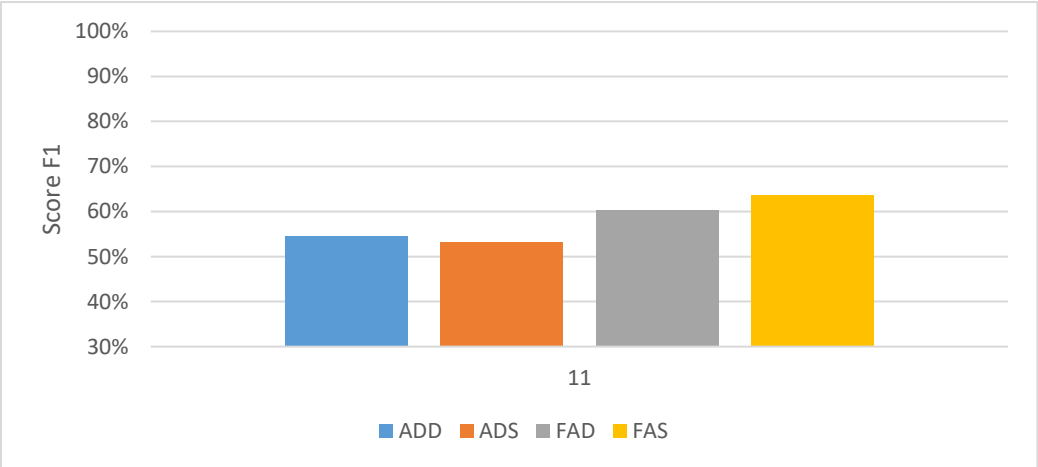

Figure 5.28 : Meilleur score F1 selon le modèle pour le client 11

Le score F1 permet de comparer facilement les modèles sur leur capacité à bien prédire la classe « annulé ». La Figure 5.28 regroupe le meilleur score F1 selon le modèle. On retrouve les bons résultats des forêts aléatoires. Le meilleur modèle est la FAS avec le meilleur score F1 de 64 % obtenu avec le regroupement 8.

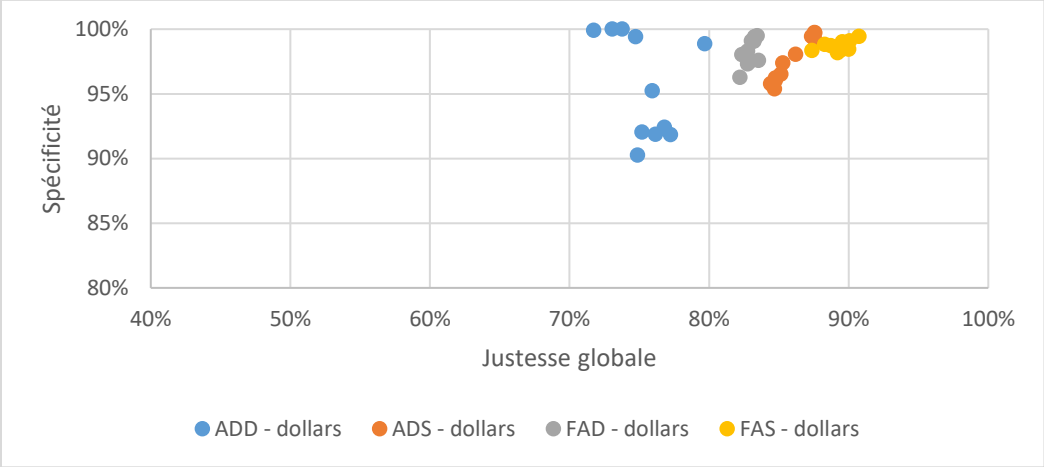

Figure 5.29 : Spécificités de la classe annulation en fonction de la justesse globale en dollars

La spécificité du client 11 en fonction de sa justesse globale est présentée à la Figure 5.29, ci-dessus. Son analyse n'est ici pas pertinente pour ce client, car la faible représentativité des devis annulés conduit à de très bons résultats.

5.3.2.2.2 Classe vente ou vente Q0

Dans cette section, nous nous intéressons à la prédiction des devis lorsque la vente est réalisée. En particulier, la précision des prédictions est évaluée ici pour la classe « VENDU » pour les modèles simples, et « Q0 » pour les modèles détaillés. On peut voir dans le graphique de la Figure 5.30 que le choix du regroupement ou du modèle n'a qu'un faible impact sur le niveau de précision.

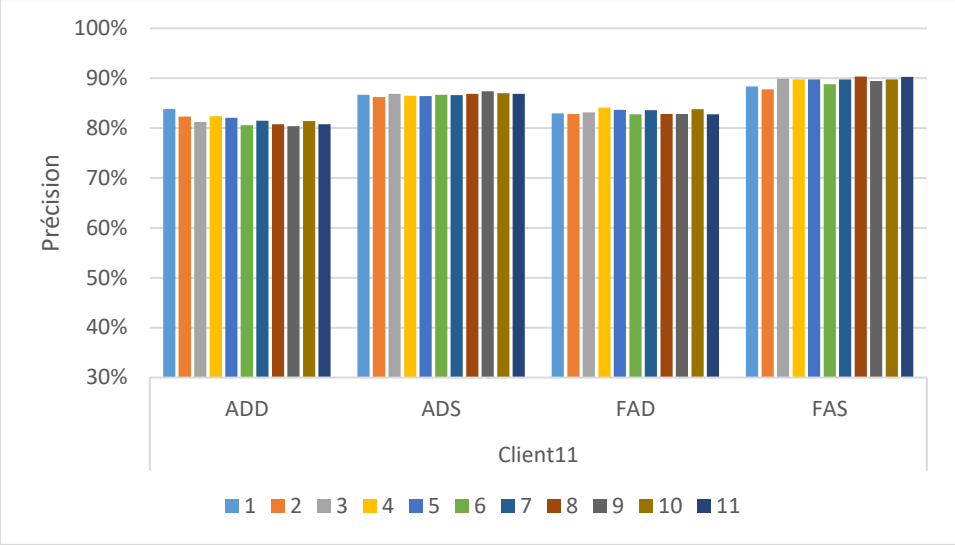

Figure 5.30 : Précision en fonction du modèle et du regroupement pour la classe vente ou Q0 (en dollars)

Les graphiques de la Figure 5.31 présentent la précision en fonction de la justesse globale à gauche (a), et de la sensibilité à droite (b). On peut voir que les devis sont généralement très bien prévus/classés pour les deux approches d'apprentissage, forêts aléatoires et arbre de décision.

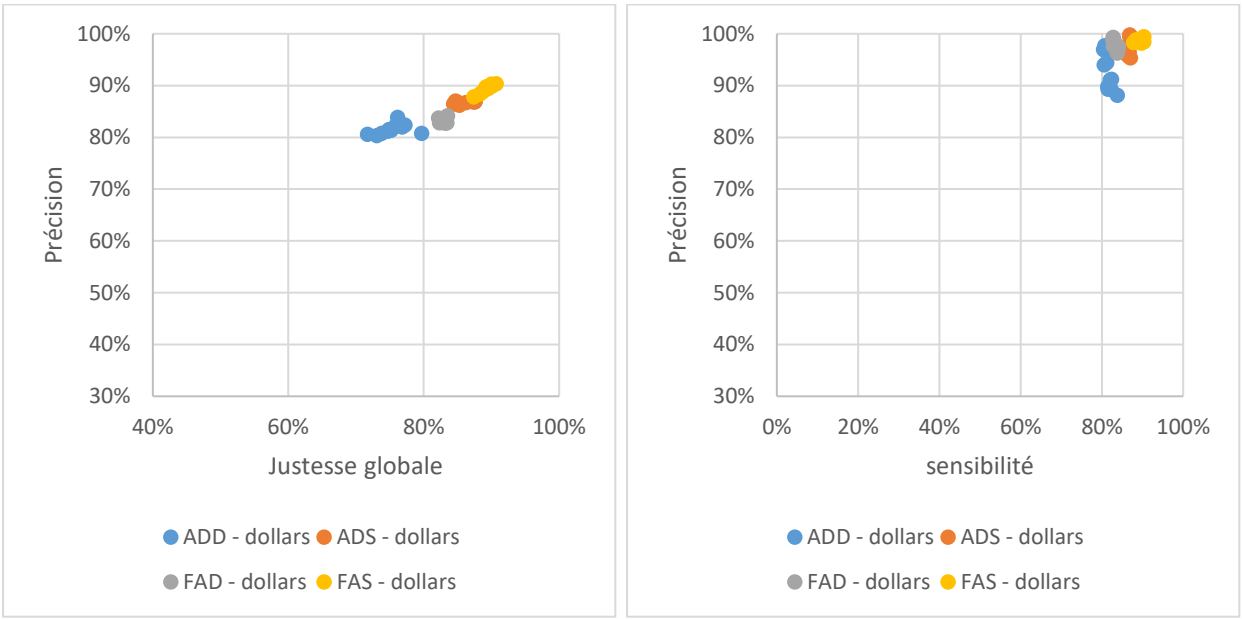

Figure 5.31 : Précision de la classe « VENDU » ou Q0 en fonction de la justesse globale (a) à gauche et en fonction de la sensibilité (b) à droite pour le client 11 (en dollars)

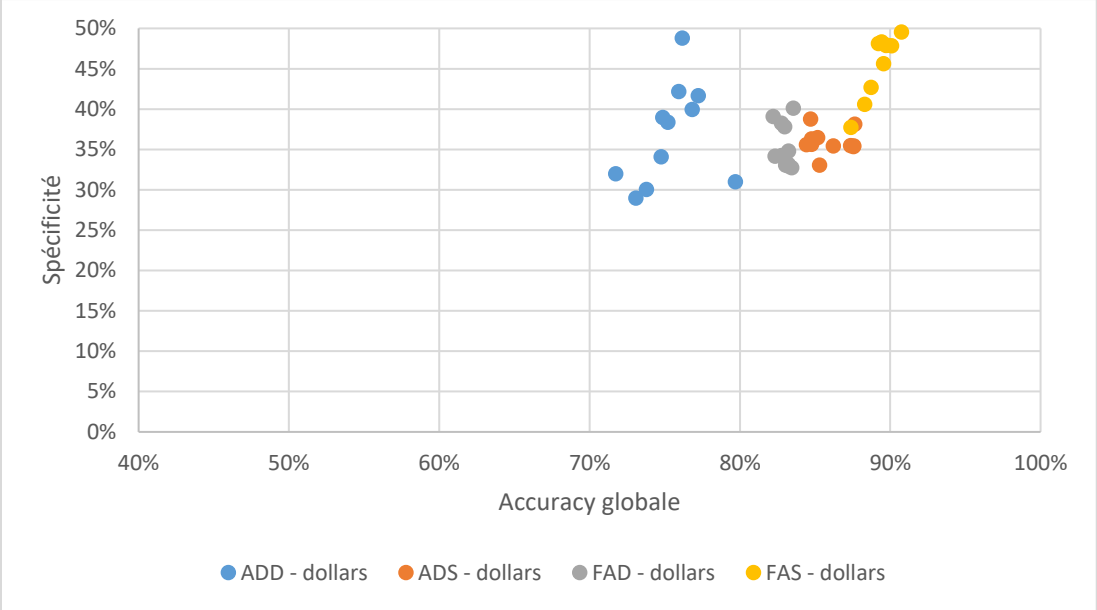

Figure 5.32 : Spécificité de la classe « VENDU» ou « Q0 » en fonction de la justesse globale en dollars

Il est intéressant de visualiser le niveau de spécificité selon le niveau de la justesse globale. La Figure 5.32 montre en effet que la spécificité faible. Cela signifie que la prédiction de l'état final des devis pour les classes « VENDU » et Q0 possède de nombreuses mauvaises classifications.

L'ensemble des résultats détaillés pour les 4 autres clients sont disponibles en annexe. Une synthèse des meilleurs résultats est présentée au Tableau 5.11

Tableau 5.11 : Meilleurs modèles pour les 5 clients étudiés

					Classe Annulation			
Client	**Meilleur Modèle**	**Quantité de données**	**Méthode de classification**	**Justesse globale**	**Score F1**	**Sensibilité**	**Précision**	**Spécificité**
11	FAS	Moyen	Produit Manhattan	91%	64%	50%	95%	99%
12	FAS	Faible	Produit Manhattan	86%	77%	74%	79%	91%
22	FAD	Faible	Seul	76%	77%	78%	77%	84%
23	FAD	Beaucoup	Argent	84%	78%	76%	81%	93%
93	FAS	Moyen	Région	92%	81%	78%	86%	96%

CHAPITRE 6 ANALYSE ET DISCUSSION

Dans ce chapitre nous discutons des résultats présentés précédemment et faisons des recommandations concernant les améliorations du processus d'enregistrement des données, ainsi que sur les travaux futurs pour l'amélioration de la qualité des prédictions.

6.1 Analyse des résultats

Dans ce chapitre, seuls les résultats pour les classes « VENDU » et « ANNULE » sont analysés. Les résultats concernant les autres classes ne sont pas suffisamment précis à cause du manque de données d'entraînement et de validation. En effet, pour les différents clients étudiés, les résultats sont similaires pour tous les modèles, quelle que soit l'approche d'apprentissage, en prenant en compte ou non le poids en dollars du devis. Les résultats suivants sont donc présentés en prenant en compte le poids des devis. Les graphiques des résultats en sans prise en compte du poids figurent néanmoins en annexe.

6.1.1 Analyse des influences des variables indépendantes

Bien que les modèles entraînés utilisent plusieurs variables indépendantes, ces dernières peuvent contribuer à différents niveaux aux calculs des prédictions. Le but de cette section est d'analyser la contribution de ces variables indépendantes.

6.1.1.1 Analyse des variables importantes

Tous les modèles étudiés ici possèdent les mêmes variables indépendantes. L'importance d'une variable dans un modèle correspond à son impact dans la prédiction finale. Il est donc important d'analyser cette composante des modèles afin de mieux comprendre l'impact des variables et isoler les plus importantes.

Pour déterminer leur impact moyen, les différentes validations croisées réalisées ont permis de déterminer l'importance de ces variables. La compilation des importances est présentée dans les Figure 6.1, Figure 6.2 et Annexe H.

6.1.1.1.1 Variable de probabilité (estimation du potentiel de réalisation des devis)

La variable de probabilité actuelle est le marqueur le plus important pour le calcul des prédictions pour tous les modèles et clients. La deuxième est la probabilité initiale. Si ces deux variables sont élevées, cela signifie que l'impact de la perception du vendeur est très fort dans les modèles. La qualité des prédictions est donc très dépendante de l'analyse subjective du vendeur.

6.1.1.1.2 Variable produit

Le contenu du devis n'est finalement pas un élément majeur dans la prédiction de son état final. Les variables de produits ont peu d'impact dans la prise de décision des modèles à l'exception du client 23 (voir Figure 6.1 et Figure 6.2). Cependant, le choix de certains produits particuliers est plus susceptible de conduire à un état final particulier. Une étude approfondie de ces produits par clients pourrait permettre d'améliorer les prédictions.

6.1.1.1.3 Analyse des autres variables

L'âge des devis est un marqueur pertinent dans l'analyse des devis. Toutefois, son impact est plus fort pour les modèles utilisant les forêts aléatoires, mais est plus limité selon les clients pour les modèles utilisant les arbres de décisions.

La variable IDprev, ou le premier CRD, est très pertinente chez certains clients pour les modèles utilisant les arbres de décisions. Pour ceux utilisant les forêts aléatoires, elle est une des 10 variables les plus importantes pour les 5 clients. Cette variable correspond à l'impact de la période de l'année sur les achats. Une modification de cette variable vers une variable cyclique sur l'année fiscale pourrait permettre d'implémenter le caractère saisonnier de certains clients.

La variable changement du CRD impacte relativement peu la décision des modèles. On note peu de changements du CRD dans les données utilisées.

Finalement, la somme totale des devis a plus de poids dans les modèles utilisant les forêts aléatoires que ceux utilisant les arbres de décisions. Cette variable est donc à conserver pour les modèles pour les premiers.

6.1.1.1.4 Comparaison entre les clients

Les variables contribuant le plus au calcul de prédiction sont similaires d'un client à l'autre. Toutefois, on peut identifier certaines spécificités. Par exemple, pour le client 23, on peut voir qu'aucune variable ne se distingue vraiment. Il est ainsi plus difficile de prédire l'état final des

devis seulement avec la probabilité donnée par le vendeur. La proportion des produits a plus d'impact.

6.1.1.1.5 Comparaison entre les modèles

Les variables de décisions pour les arbres de décisions simples et détaillés sont très similaires (Figure 6.1, Figure 6.2 et Annexe H). Les figures montrent l'importance des 10 variables les plus influentes pour chacun des clients pour les modèles ADD et FAD. Les boîtes à moustache sont obtenues par validation croisée et le calcul de l'importance à chaque itération.

L'impact des variables n'est pas équilibré pour les modèles ADD. Par exemple, la Figure 6.1 montre des variations, entre la médiane de chacune des importances des différentes variables, entre 2 et 35 pour le client 11. Les modèles avec forêts aléatoires sont plus équilibrés entre les variables. Pour les modèles basés sur les arbres de décisions, les variations sont plus fortes. À la Figure 6.2, on peut voir que les variations des médianes pour le client 11 sont entre 2 et 20. Le manque d'efficacité des modèles types ADD et ADS peut être en partie expliqué par la dominance de certaines variables. L'équilibre d'importance des variables pour les modèles types FAS et FAS permet de meilleurs résultats.

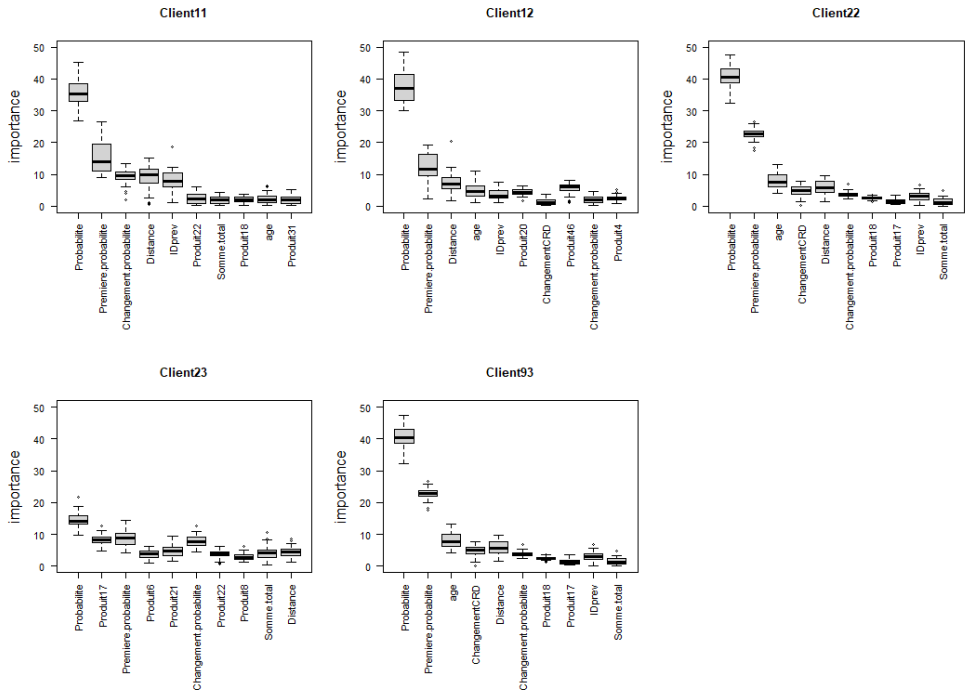

Figure 6.1 : Importance des variables pour les arbres de décision détaillés pour les 5 clients

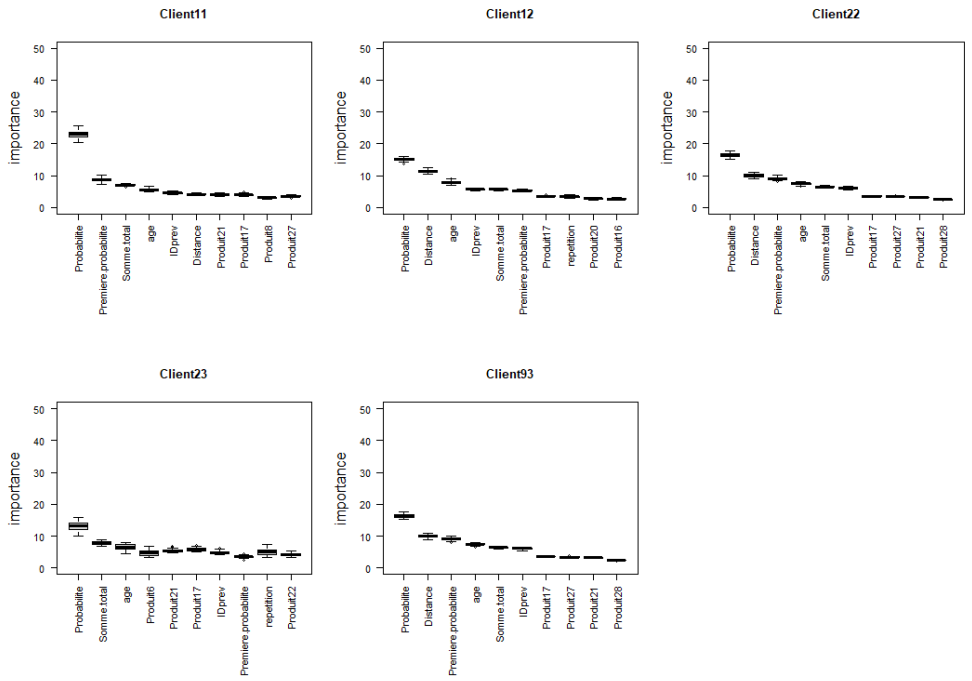

Figure 6.2 : Importance des variables pour les forêts aléatoires simples pour les 5 clients

6.1.1.2 Justesse globale

Pour le client 11, le choix d'un regroupement a un impact relativement faible sur les résultats, quel que soit le modèle choisi.

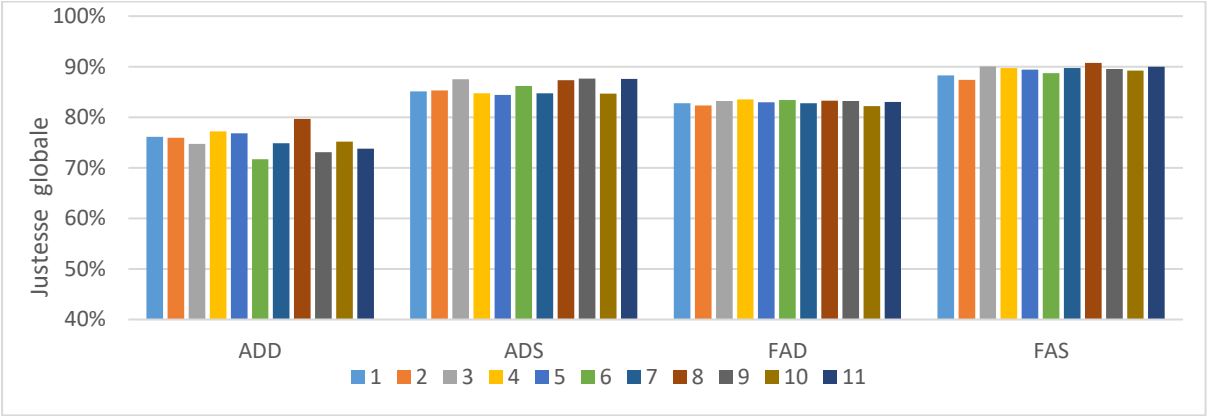

Figure 6.3 : Justesse globale en fonction du regroupement et des modèles pour le client 11

Les impacts des différents regroupements sont plus forts pour les modèles utilisant les arbres de décisions que pour les modèles utilisant les forêts aléatoires. Pour le modèle ADD (arbre de décision avec modélisation détaillée de l'état final), il y a une différence entre le meilleur et le

moins bon modèle de l'ordre de 8 %, alors que pour le modèle FAD (forêts aléatoires avec modélisation détaillée de l'état final) l'impact est négligeable. La Figure 6.4 montre l'impact négligeable avec la classification produit Manhattan.

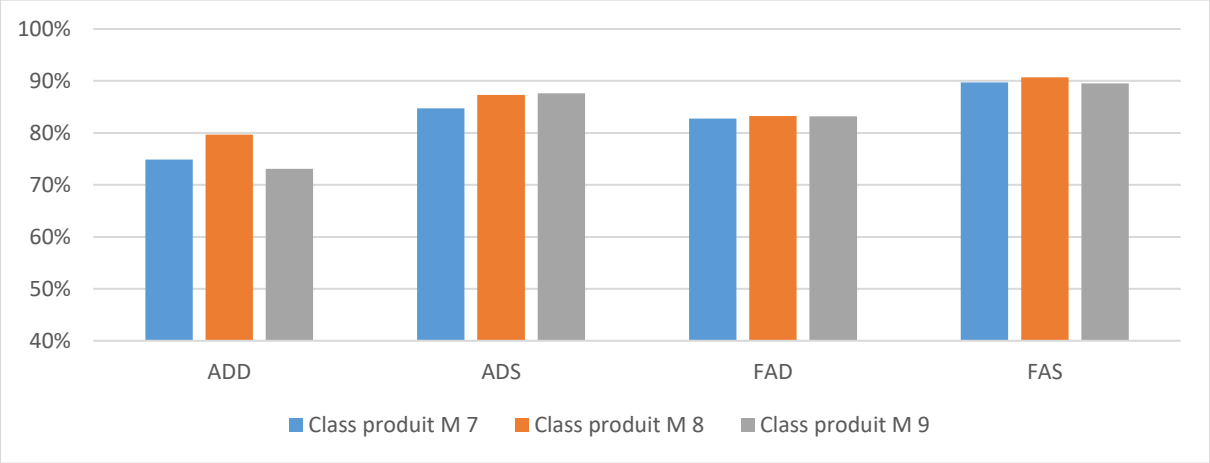

Figure 6.4 : Justesse globale en fonction des modèles pour la classification produit Manhattan (client 11)

Pour les autres clients (12, 22 et 93), les regroupements impactent plus la justesse. On peut donc dire pour ces clients que les données d'entraînement ont plus d'impact. Il faut bien choisir leur regroupement pour obtenir la meilleure classification globale.

Pour le client 22, comme indiqué sur la Figure 6.5, les regroupements argent ou régions donnent les meilleurs résultats pour les modèles FAS. Les modèles utilisant ADS et FAS présentent une variation relativement faible entre les classifications. Ceci est dû à la présence de seulement deux classes. Les prédictions globales ont moins de chance de se tromper.

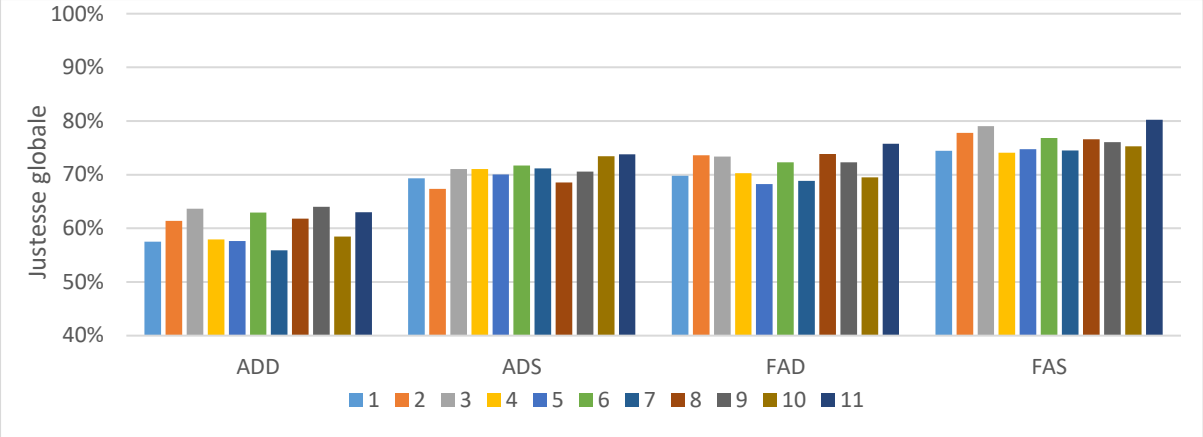

Figure 6.5 : Justesse globale en fonction du regroupement et des modèles pour le client 22

L'augmentation du nombre de données n'améliore pas les résultats de la justesse. La Figure 6.5 permet de voir que l'augmentation du nombre de données d'entraînement n'augmente pas la justesse. Cet indicateur ne semble pas être influencé par le volume de données. La meilleure justesse globale pour le client 22 est visible avec le regroupement « Seul » et les modèles FAS (forêts aléatoires et modélisation simple de l'état final). Les écarts maximums pour les autres modèles sont au plus de 8 % (Figure 6.6). Le choix de la modélisation de l'état final (simple ou détaillée, cf. 5.2.6.3) a ainsi plus d'impact sur la justesse que la quantité de données traitées pour les clients 11, 12 et 93. Il vaut mieux choisir pour ces clients des classifications simples si on cherche à avoir la meilleure justesse globale. Pour les clients 22 et 23, il vaudra mieux privilégier l'approche d'apprentissage par forêts aléatoires.

Les modèles utilisant une modélisation simple de l'état final prévu des devis sont meilleurs que les modèles avec modélisation détaillée, car ils possèdent moins de classes. La moins bonne performance des clients 22 et 23 peut être expliquée par le nombre plus important de devis annulés (Figure 5.21).

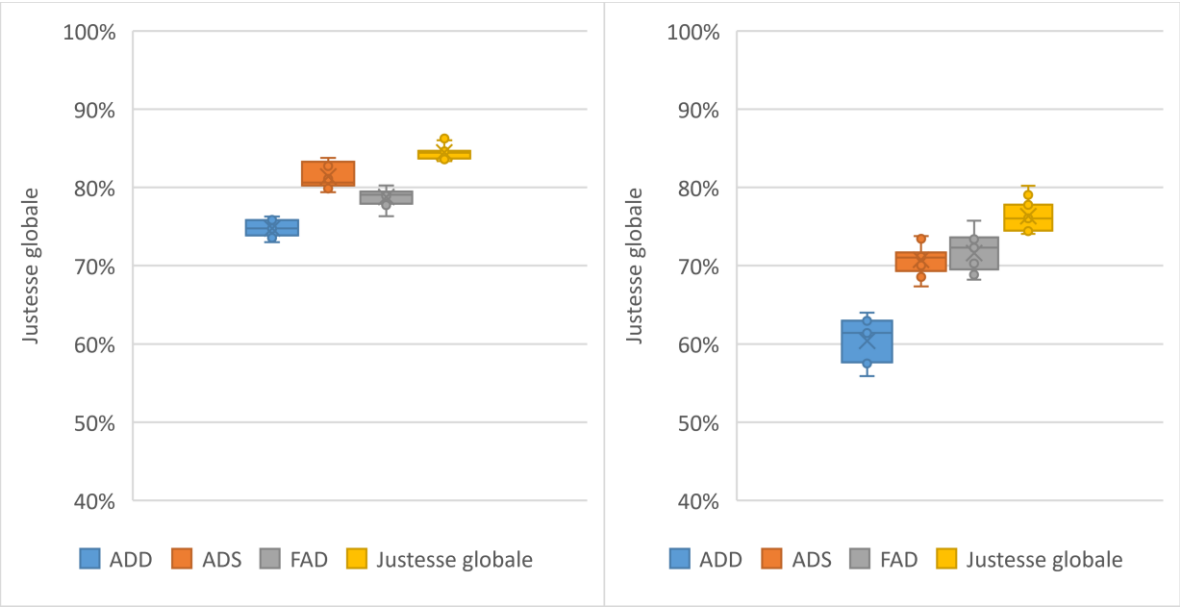

Figure 6.6 : Influence du regroupement sur la justesse globale (Client 12 à gauche et Client 22 à gauche)

La quantité de données n'est donc pas le seul élément à prendre en compte pour améliorer les modèles. Pour chaque client, il y a une méthode optimale de classification qui lui est propre comme on peut le voir dans le Tableau 6.1. On peut lire dans le tableau la meilleure justesse par modélisation et approche. La colonne quantité de données renvoie aux volumes de données relatifs dans l'étude. Les deux colonnes qui suivent donnent le regroupement et son numéro ou la meilleure justesse est obtenue. Les lignes en gros correspondent au meilleur approche d'apprentissage. Les forêts aléatoires sont toujours meilleures, quel que soit le client. Cette méthode est donc à privilégier pour avoir une justesse globale meilleure.

Tableau 6.1 : Meilleurs modèles selon la justesse globale (en dollars)

Client	Modèle	Quantité de données	Méthode de regroupement	Meilleur modèle justesse globale	Justesse globale
11	ADD	Moyen	Produit M	8	80%
	FAD	**Beaucoup**	**Produit E**	**4**	**84%**
	ADS	Faible	Produit M	9	88%
	FAS	**Moyen**	**Produit M**	**8**	**91%**
12	ADD	Faible	Produit E	6	76%
	FAD	**Moyen**	**Produit M**	**8**	**86%**
	ADS	Moyen	Région	10	84%
	FAS	**Faible**	**Produit M**	**9**	**86%**
22	ADD	Faible	Produit M	9	64%
	FAD	**Faible**	**Seul**	**11**	**76%**
	ADS	Faible	Seul	11	74%
	FAS	**Faible**	**Seul**	**11**	**80%**
23	ADD	Faible	Produit M	9	81%
	FAD	**Faible**	**Produit E**	**6**	**84%**
	ADS	Faible	Argent	3	81%
	FAS	**Faible**	**Produit E**	**6**	**87%**
93	ADD	Moyen	Argent	2	74%
	FAD	**Faible**	**Région**	**10**	**79%**
	ADS	Moyen	Produit M	8	84%
	FAS	**Faible**	**Région**	**10**	**92%**

6.1.1.3 Analyse de la classe « ANNULÉ »

La classe « ANNULÉ » est la classe la plus importante pour l'entreprise partenaire. Elle souhaite en effet pouvoir identifier et anticiper les devis qui ne seront pas transformés en vente, afin d'éviter une surproduction. Pour analyser la classe « ANNULÉ », nous avons étudié les indicateurs suivants : la précision, la sensibilité, le score F1 et la spécificité.

6.1.1.3.1 La précision et justesse globale

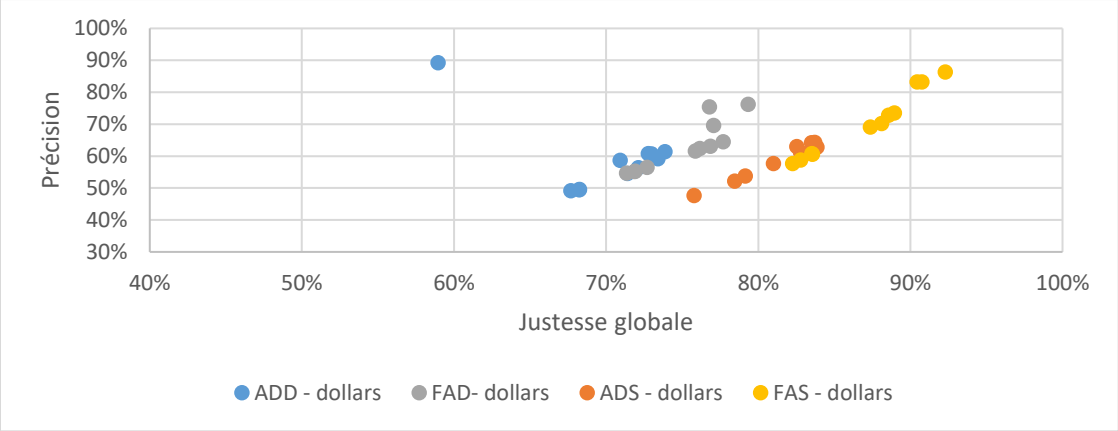

Figure 6.7 : Précision en fonction de la justesse globale pour le client 93

La justesse et la précision sont parfois corrélées, l'optimisation de ces deux paramètres peut se faire en parallèle. Pour les clients 22, 23 et 93, plus la justesse globale augmente plus la précision augmente. Pour le client 22, qui possède un nombre conséquent de devis « ANNULÉ » (Figure 5.21), ça signifie que l'on peut obtenir une bonne précision sans rogner sur une bonne justesse globale. On retrouve les deux groupes correspondant au choix de la modélisation de l'état final en regardant la précision en fonction de la justesse : les modèles utilisant la classification simple et ceux utilisant la classification détaillée. Pour le client 93, les modèles FAS dominent les autres.

On peut noter que le choix du regroupement impacte fortement les résultats. Pour le modèle FAS, la Figure 6.7 montre une différence de précision de 34 % entre le pire modèle et le meilleur. Le meilleur modèle (86 % de précision avec 92 % de justesse) est obtenu pour le client 93 pour la classification par région. Pour les autres clients, les résultats se situent entre 50 et 95 %. Cela signifie que les annulations annoncées dans les modèles vont être réellement annulées. Il y a peu de faux positifs, donc peu de devis prédits annulés qui seront finalement réalisés.

6.1.1.3.2 Précision et sensibilité

Au-delà de la précision, il faut savoir si tous les devis annulés ont bien été identifiés comme tels. En plus de la certitude de n'avoir que des devis annulés dans la classe « ANNULÉ », il faut aussi détecter l'ensemble de tous les devis ayant l'état final « ANNULÉ ». Cette caractéristique est représentée par la sensibilité (Figure 6.8).

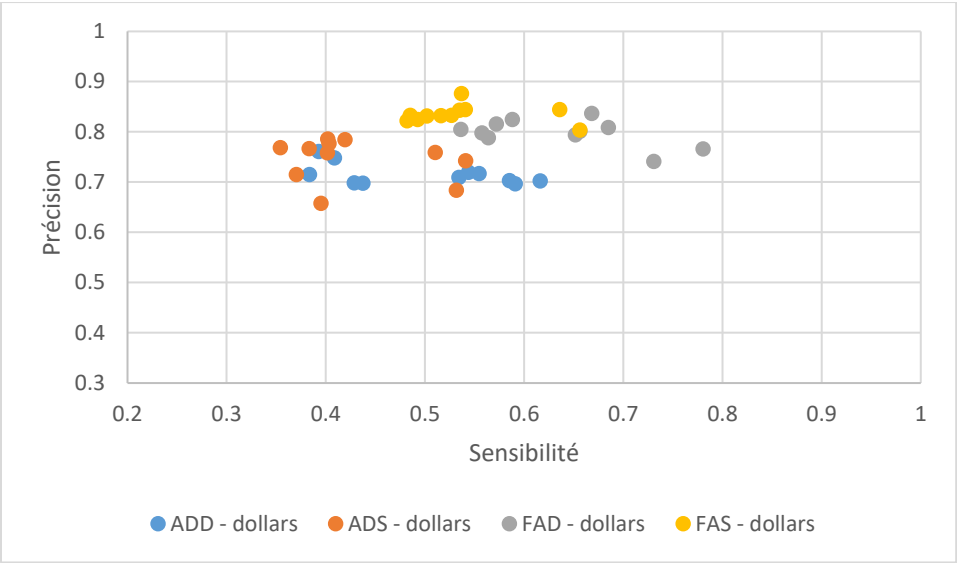

Figure 6.8 : Sensibilité en fonction de la précision pour le client 23 (annulation)

La sensibilité est très variable. La qualité des données d'entrée est importante. La meilleure sensibilité évolue selon les clients de 53 % à 91 %. Un choix de regroupement doit donc être réalisé pour obtenir aussi bien une précision et une sensibilité acceptable. Les modèles FAD sont les meilleurs pour identifier le maximum de devis annulés. La Figure 6.9 montre la meilleure sensibilité selon le type de modélisation pour les cinq clients. Malgré la présence de 6 classes différentes, FAD réussit à prédire « ANNULÉ » jusqu'à 90 % des devis ayant pour état final la classe « ANNULÉ ». Le client 11 est plus difficile à prédire, car il présente peu de devis annulés (Figure 5.21).

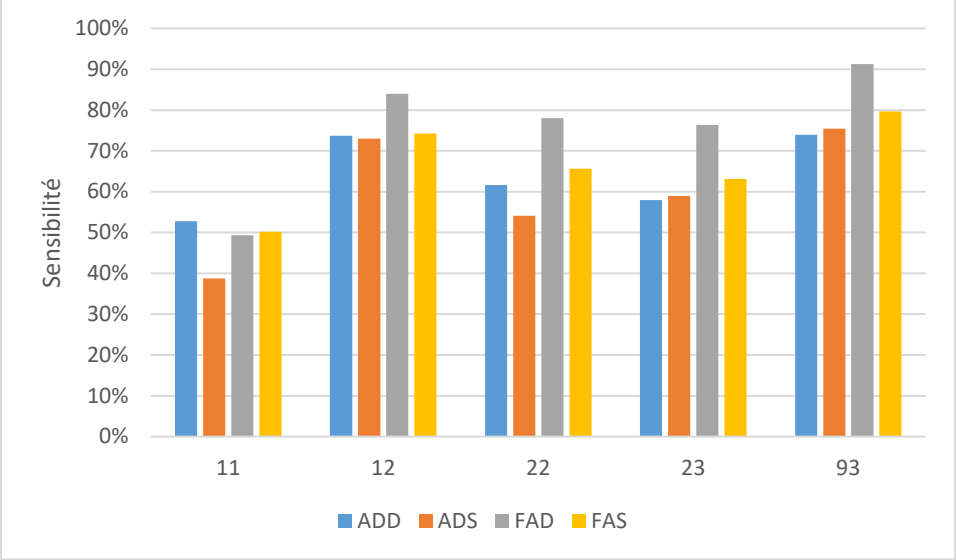

Figure 6.9 : Meilleures sensibilités (annulation) pour chaque type de modèle et client

6.1.1.3.3 Le score F1

Après l'étude de la précision et de la sensibilité, le score F1 va permettre de trouver un compromis entre le taux de faux positifs et de faux négatif.

Pour identifier le meilleur modèle selon les deux précédents indicateurs, on utilise le score F1. Ce dernier rassembler les deux indicateurs et donne un avis d'ensemble. Le score est, selon les clients, fortement influencés par le regroupement choisi comme on peut le voir à la Figure 6.10.

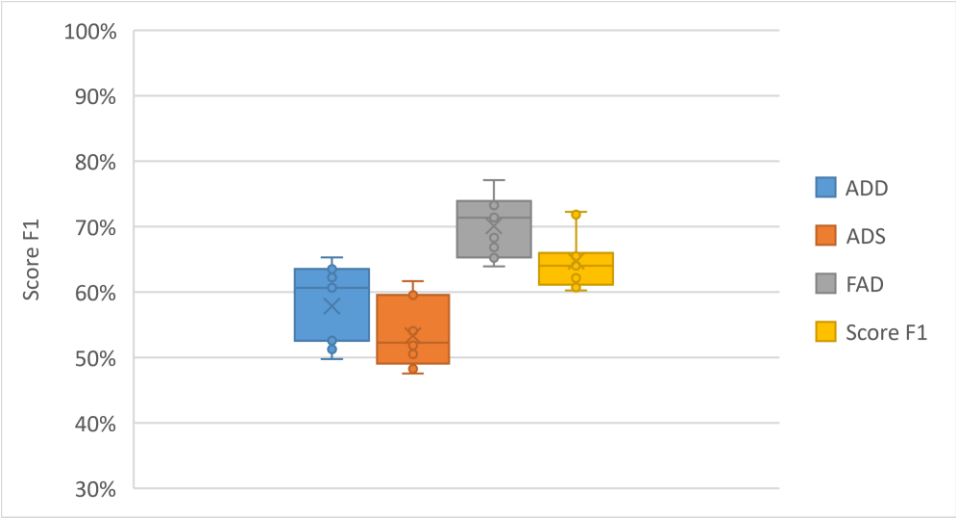

Figure 6.10 : Influence du regroupement sur le score F1 (Client 22).

Tous les couples modèles-clients ne sont pas influencés aussi fortement. Pour le client 22, on peut noter que, selon le regroupement choisi, pour l'arbre de décision simple, le score F1 peut être de 48 % ou de 62 %. Il faut donc prendre le temps d'identifier le meilleur regroupement pour chaque client afin d'avoir une qualité de prédiction satisfaisante.

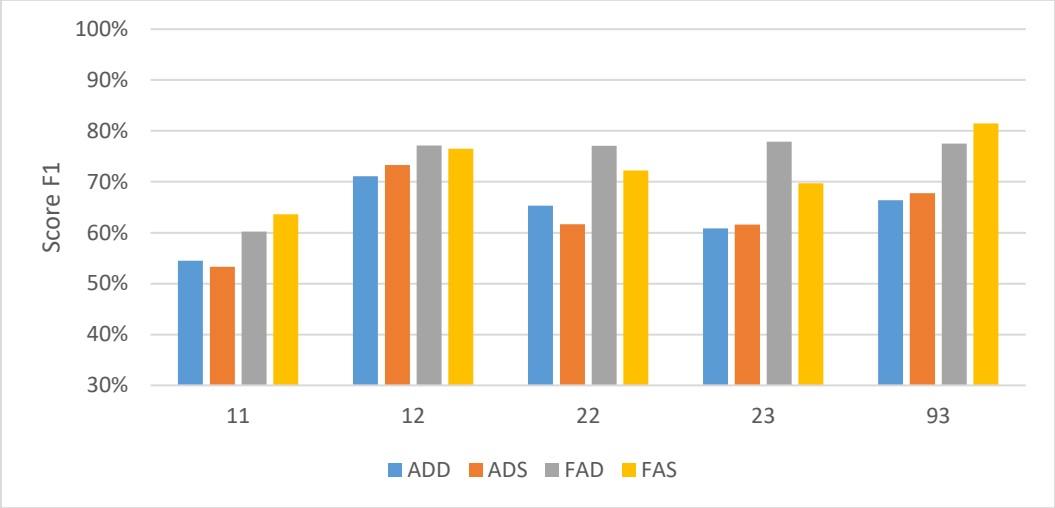

Figure 6.11 : Meilleure score F1 (annulation) pour chaque type de modèle et client

Les clients 11 et 93 obtiennent leur meilleur résultat avec les forêts aléatoires simples. Nous avons utilisé le score F1 de la classe « ANNULÉ » pour déterminer les modèles recommandés à l'entreprise. Les résultats mettent en avant le fait que les forêts aléatoires permettent de meilleures prédictions pour tous les clients. Le tableau 21 montre le type de regroupement qui permet d'obtenir ces résultats.

6.1.1.3.4 Les meilleurs modèles pour détecter les annulations

Le Tableau 6.2 permet d'identifier quel modèle doit être privilégié pour chaque client pour détecter un maximum de devis annulés et recevoir peu de fausses alertes.

Les différents modèles sont efficaces pour attribuer la classe annulation aux devis réellement annulés cependant, beaucoup de devis annulés ne sont pas identifiés.

Le meilleur modèle de détection change selon le client. Les meilleurs modèles de prédictions d'annulation apparaissent, en gras, dans le Tableau 6.2. Pour le client 11, on obtient les meilleurs résultats avec les modèles FAS regroupement 8. Pour ce client, on arrive à détecter la moitié des devis ayant l'état final « ANNULÉ » avec une précision de 95 %. La justesse globale n'est pas

exceptionnelle aux vues de la prévalence (Figure 5.21) de la classe « VENDU ». Ce mauvais score est donc justifié par un manque de devis « ANNULE ». Pour le client 22, la prévalence est beaucoup plus équilibrée (Figure 5.21). Les résultats sont très bons. On obtient les meilleurs résultats avec FAD et le regroupement 11, c'est-à-dire avec seulement ses devis. On obtient pour ce client, 78 % de sensibilité et 77 % de précision. La justesse est cependant un peu faible avec 76 % de bonne classification. Le client 22 est un bon exemple que la prédiction de l'état final des devis est possible.

Les meilleurs résultats sont obtenus par le client 93 avec le modèle FAS et le regroupement 10. On obtient pour ce client, 78 % de sensibilité et 86 % de précision. La justesse est excellente avec 92 % de bonne classification.

Tableau 6.2 : Résumé des meilleurs scores F1 par modèle et par clients

Client	Modèles	Quantité de données	Méthode de classification client	Meilleur modèle selon le score F1 (annulation)	Score F1	Sensibilité	Précision	Justesse globale	Spécificité
11	ADD	Beaucoup	Argent	1	55%	53%	60%	76%	92%
	FAD	Beaucoup	Produit euclidien	4	60%	49%	80%	84%	98%
	ADS	Faible	Produit Manhattan	9	53%	38%	93%	88%	99%
	FAS	**Moyen**	**Produit Manhattan**	**8**	**64%**	**50%**	**95%**	**91%**	**99%**
12	ADD	Faible	Produit euclidien	6	71%	72%	72%	76%	87%
	FAD	Moyen	Produit Manhattan	8	77%	84%	72%	80%	86%
	ADS	Moyen	Région	10	73%	73%	74%	84%	88%
	FAS	**Faible**	**Produit Manhattan**	**9**	**77%**	**74%**	**79%**	**86%**	**91%**
22	ADD	Faible	Argent	3	65%	62%	70%	64%	83%
	FAD	**Faible**	**Seul**	**11**	**77%**	**78%**	**77%**	**76%**	**84%**
	ADS	Faible	Seul	11	62%	54%	74%	74%	87%
	FAS	Faible	Seul	11	72%	64%	84%	80%	92%
23	ADD	Beaucoup	Argent	1	61%	56%	67%	74%	90%
	FAD	**Beaucoup**	**Argent**	**1**	**78%**	**76%**	**81%**	**84%**	**93%**
	ADS	Moyen	Produit Manhattan	8	62%	59%	65%	81%	89%
	FAS	Faible	Produit euclidien	6	70%	61%	83%	87%	95%
93	ADD	Moyen	Argent	2	66%	74%	61%	74%	86%
	FAD	Moyen	Région	10	78%	81%	76%	79%	93%
	ADS	Moyen	Argent	2	68%	75%	62%	84%	86%
	FAS	**Moyen**	**Région**	**10**	**81%**	**78%**	**86%**	**92%**	**96%**

6.1.1.4 Analyse classe « Q0 » - vendu sur le bon quart et autres classes

Dans cette partie, nous analyserons la classe Q0 sur les modèles ADD et FAD. L'objectif est d'analyser une autre classe que la classe « ANNULÉ » pour les modèles détaillés. Le modèle FAD étant le meilleur pour 3 clients sur les 5, il est important d'analyser la classe dominante.

6.1.1.4.1 Précision et sensibilité pour Q0

La précision des prédictions des ventes dans le bon quart issu des modèles détaillés est correcte pour tous les clients (autour de 80 % pour le client 11), sauf pour le client 22 (Figure 5.21). Pour ce dernier, la plus grande proportion de devis annulés peut expliquer pourquoi la classe Q0 est moins précise. Il est plus facile en effet d'obtenir 80 % de précision lorsque 80 % ou plus des devis sont Q0. Les bons résultats des autres clients peuvent être expliqués par le nombre plus important de devis vendus utiles pour l'entraînement des modèles.

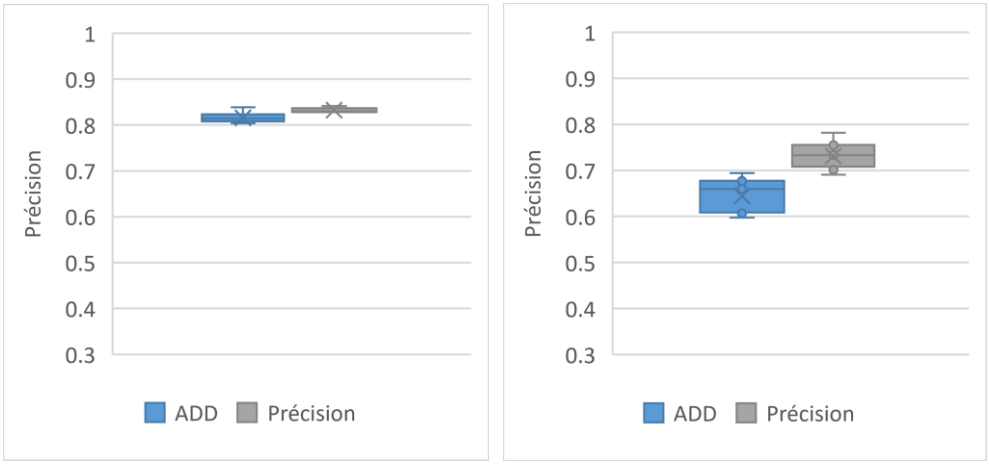

Figure 6.12 : Précision selon les modèles pour le client 11 (à gauche) et 22 (à droite)

Concernant la sensibilité, les médianes pour tous les clients sont au-dessus de 80 %, avec peu de variations au sein d'une modélisation et approche. Les regroupements ont donc peu d'influence dans l'entraînement des modèles. L'influence vient plus du type d'approche. Le choix des modèles FAD devient important, car il a une meilleure sensibilité et une moins grande variation (5.3.1)

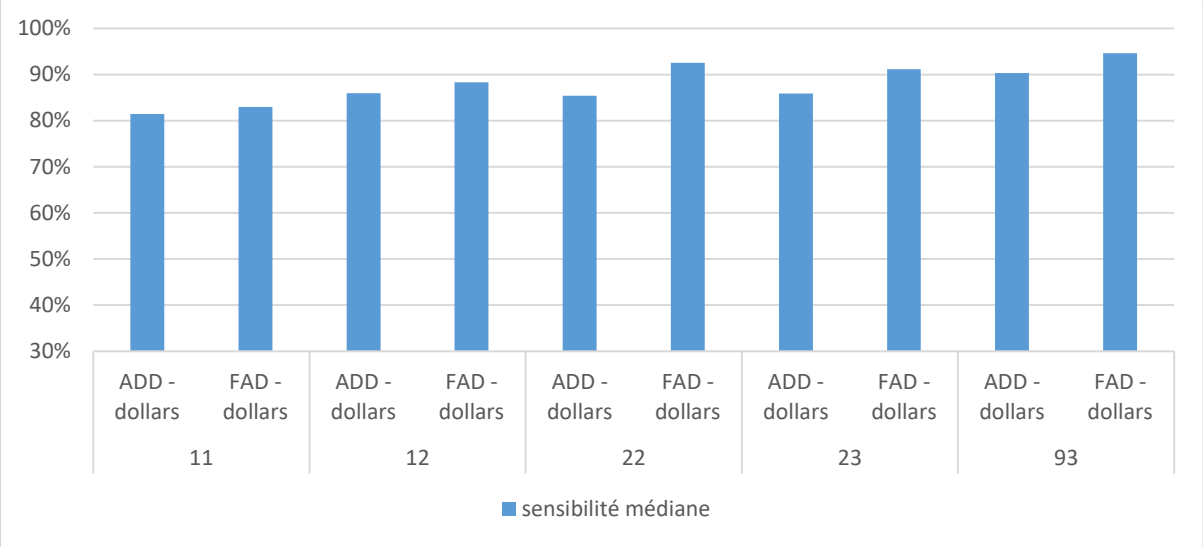

Figure 6.13 : Sensibilité médiane selon le modèle pour les cinq clients

6.1.1.4.2 Autres classes

Les autres classes ne peuvent pas être analysées correctement à cause du manque de données.

En effet, on peut voir dans le Tableau 6.3 : Moyenne et écart type du score F1 (pour Q-1) selon le regroupement (Client 22), la moyenne et écart type du score F1 pour la prévision un quart en avance (Q-1) (cf. 5.2.3). Dans le tableau on peut voir des moyennes très faibles et des écarts types élevés. Ces valeurs montrent une grande variabilité des résultats. Le manque de données ne permet pas de conclure sur une analyse. On retrouve ces résultats pour les 5 clients et pour les différentes classes tel que Q-1, Q+1, Q+2 et Q+3.

Tableau 6.3 : Moyenne et écart type du score F1 (pour Q-1) selon le regroupement (Client 22)

Regroupement	1	2	3	4	5	6	7	8	9	10	11
ADD											
Moyenne de score F1	19%	26%	30%	13%	13%	21%	16%	21%	24%	26%	31%
Écart type de score F1	13%	9%	19%	8%	8%	11%	7%	11%	7%	10%	14%
FAD											
Moyenne de score F1	52%	56%	62%	51%	51%	60%	50%	61%	61%	55%	61%
Écart type de score F1	26%	23%	20%	24%	25%	19%	25%	19%	19%	26%	19%

6.1.1.4.3 Conclusion

Les modèles de prédiction détaillés étudiés rencontrent généralement des difficultés à prévoir de nombreuses classes. Autrement dit, aucun d'eux ne peut prévoir précisément le retard des devis.

Cependant, les bons résultats de la prédiction des devis vendus, soit la classe Q0, démontrent que les modèles testés peuvent correctement prédire les ventes.

Cependant, ces modèles ont beaucoup de mal à classer les devis des autres classes. La matrice de confusion présentée au Tableau 6.4 présente dans quelles classes les devis sont prédits pour le client 23 et un modèle spécifique. Chaque case est divisée par le total de sa ligne. Ainsi, 93 % des dollars qui ont été prédits en Q0 ont bien été vendus. Pour les ventes réalisées avec un quart de retard, soit la classe Q1, 47 % avaient été prédits en Q0 et 43 % en Q4. On peut aussi noter le faible pourcentage de Q0 prédit en annulation (Q4). Les devis en retard de plus de 2 trimestres sont souvent classés dans la classe « ANNULE ». Ceci peut être dû à la ressemblance entre ces devis. De même pour les devis en avance de 1 trimestre ou en retard d'un trimestre avec la classe Q0.

Tableau 6.4 : Matrice de confusion du client 23 pour le modèle ADD regroupement 5

		Prédit					
		Q-1	Q0	Q1	Q2	Q3	Q4
Réel	Q-1	0%	95%	0%	0%	0%	5%
	Q0	0%	93%	1%	0%	0%	5%
	Q1	0%	47%	10%	0%	0%	43%
	Q2	0%	12%	8%	3%	0%	77%
	Q3	0%	0%	0%	0%	0%	100%
	Q4	0%	25%	2%	0%	0%	72%

6.1.2 Limites et améliorations potentielles

Malgré les bons résultats de la prédiction des devis vendus, de nombreuses pistes sont possibles afin d'améliorer ces modèles.

6.1.2.1 Le nombre de clients

Actuellement, les modèles ont été testés avec 5 clients. L'ajout d'une quantité plus importante de données de devis permettrait d'analyser un plus grand nombre de clients. Il faudrait, par la suite, tester les modèles sur d'autres clients et d'autres regroupements de clients.

6.1.2.2 Le choix des variables

Les variables de décisions choisies ne sont pas parfaites. Comme nous avons pu le voir dans leur l'analyse (6.1.1.1), certaines variables contribuent peu, voire pas du tout, au calcul des prédictions.

D'autres variables pourraient donc être ajoutées sur une plus grande temporalité telles que :

- Les dépenses en ventes directes sur l'année ou le quart précédents. Elles interagissent avec les devis du fait des achats effectués en parallèle. Les ventes directes pourraient avoir un impact sur l'évolution des devis et leur modification.
- La saisonnalité des achats de certains clients pourrait avoir un impact sur les ventes. On peut citer l'exemple d'une indication de saisonnalité pour les clients 22 et 125 (Figure 6.14). Cet élément mis en œuvre à cause du manque de données.
- Un score de réalisation des quarts précédents. Le comportement du client sur les quarts précédents pourrait avoir un impact sur les ventes futures et l'enregistrement de ce score et l'application pour la modélisation seraient pertinents.

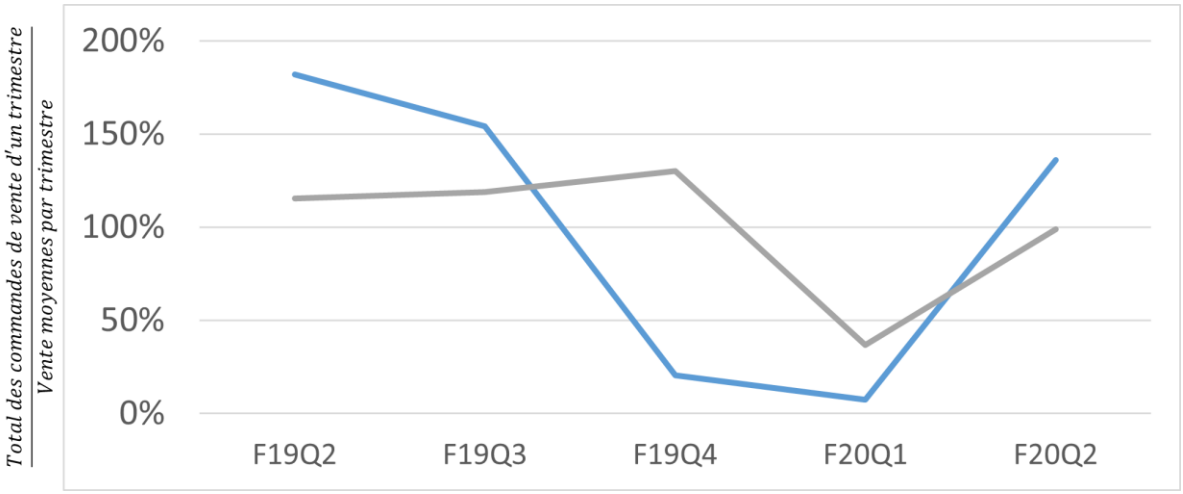

Figure 6.14 : Évolution des ordres de vente par quart (Client 22 en gris et 125 en gris)

6.1.2.3 Ventes directes

Dans les modèles testés, les ventes directes ne sont pas prises en compte. Cependant elles peuvent représenter une part importante des achats de certains clients comme indiqué dans la partie 4.2.3.2. Il serait pertinent de prendre en compte ces achats dans une prochaine étude.

6.1.2.4 Réparation des données

La réparation des données effectuée dans la partie 5.1 est une limite de la modélisation. L'impossibilité de « recoller » les devis initiaux nous a obligé à appliquer une correction du processus d'enregistrement de la base de données. Or cette correction a pu écarter des informations qui n'auraient donc pas été prises en compte dans notre étude. Le processus d'enregistrement des devis « PRÉVISION - VENTE FERME » a pu ainsi causer des erreurs dans lors modification en « ORDRE DE VENTE ». En effet, comme nous l'avons vu précédemment, leurs dates réelles de vente n'étant pas connues, nous avions décidé de leur attribuer la date d'arrivée dans cet état comme date de vente. Un biais a ainsi pu être introduit à ce moment.

6.1.2.5 Base de données d'entraînement

Le regroupement des clients a été limité à 25. Dans de futurs travaux, nous pourrions élargir ces regroupements. Ces derniers seraient possiblement plus pertinents et permettraient d'obtenir de meilleurs résultats. De nouveaux tests sont à prévoir pour affiner l'entraînement des modèles. On pourra aussi élargir le champ des données sur davantage de devis en remontant plus loin dans l'historique des données. Nos modèles ont été entraînés sur un maximum de 5 trimestres. Ce nombre pourrait ainsi être augmenté. Il faudra cependant prêter attention aux trimestres de 2020, très impactés par la pandémie.

6.2 Recommandations

6.2.1 Les devis et processus

Dans cette section, nous présentons les améliorations qui permettraient à l'entreprise partenaire de réaliser un meilleur suivi des devis et d'améliorer leur processus de prédiction.

6.2.1.1 Le cas des « PRÉVISION – FERME GAGNE »

De nombreux devis passent par l'état « PRÉVISION – FERME GAGNE ». Cet état pose de nombreux problèmes dans le suivi des devis dans le temps. Une grande partie du projet s'est concentrée sur des tentatives de « recollage » des devis et des ordres de vente. Ce « recollage » a été impossible du fait d'un nombre trop important d'obstacles. Une solution à ce problème serait d'attribuer un même identifiant à ce type de devis. Nous pourrions ainsi faire un suivi complet et mieux comprendre l'arrivée de certaines commandes en mode direct.

6.2.1.2 Suppression des devis inutiles dans le temps

Certains devis restent plusieurs mois dans les bases de données bien qu'ils aient atteint un état final. Ce type de devis complique l'exploitation des données. Il faudrait donc systématiser l'utilisation de la base de données par les vendeurs, et en particulier, standardiser les processus de suppression des devis inutiles en termes de suivi, dont l'état final a été atteint.

6.2.1.3 Stabilité des clients

La modification de la dénomination de clients au sein d'un même devis peut arriver. Ce type de changement est gênant dans la prédiction des ventes de ces clients. Pour rester utilisables, les devis devraient changer de nom lors d'un tel changement, et le devis initial devrait évoluer en annulation avec un indicateur « ANNULER – CHANGEMENT CLIENT ». Ce suivi permettrait une plus grande précision des données et donc des modèles.

6.2.1.4 Annulation et vente

Certains devis suivent une séquence illogique d'états. Ils commencent par un cycle de vie classique, arrivent dans l'état annulé, mais sont finalement vendus peu de temps après. Le devis vendu peut présenter des produits totalement différents du devis initial. Ce type d'incohérence cause des difficultés de compréhension de leurs cycles de vie. L'attribution de l'état annulé à un devis devrait ainsi empêcher toute modification par la suite.

6.2.1.5 Devis classique VS Devis en vrac

Les deux processus de ventes cohabitent au sein de l'entreprise. Il n'y a aujourd'hui aucune façon claire de différencier les deux types de devis. Une numération différenciée permettrait une

meilleure séparation des clients et des devis. En effet, certains clients fonctionnent seulement avec l'un ou l'autre des formats. Il serait plus facile de choisir des clients à analyser en utilisant le cycle de vis des devis, et d'autres en utilisant des séries chronologiques.

6.2.2 Clients et vendeur

6.2.2.1 Stabilité de la hiérarchie des clients

Lors du projet, nous avons ainsi appris que certaines régions avaient changé de nom. Ces changements peuvent modifier en profondeur la hiérarchie des clients et compliquer énormément l'historique des regroupements des clients. La création d'une cartographie des liens clients, et d'un historique de ces liens permettrait de créer des bases de données utiles plus volumineuses.

6.2.2.2 Lien vendeur – client

Le lien entre les vendeurs et les clients est difficile à évaluer. Ce lien est très subjectif et dépend beaucoup de la région de travail. La probabilité utilisée comme variable indépendante est le seul indicateur de confiance que donne le vendeur. Celui-ci dépend fortement de son analyse subjective. Une évaluation de la capacité du vendeur à émettre une bonne évaluation permettrait certainement une meilleure compréhension de certaines surévaluations ou sous-évaluations des ventes. Un guide standard de cette évaluation pourrait aussi uniformiser le caractère subjectif de l'évaluation de chaque vendeur. Une autre possibilité serait d'ajouter des critères objectifs à renseigner en plus de la probabilité pour un meilleur contrôle de l'information.

6.2.2.3 Utilisation pratique des modèles

6.2.2.3.1 Utilisation des modèles détaillés

Les modèles détaillés actuels ne donnent pas toujours des résultats concluants pour la prédiction de l'état final. Cependant, ils peuvent être utilisés comme indicateur. La probabilité des différentes classes permet de mieux appréhender l'évolution du devis dans un sens ou un autre.

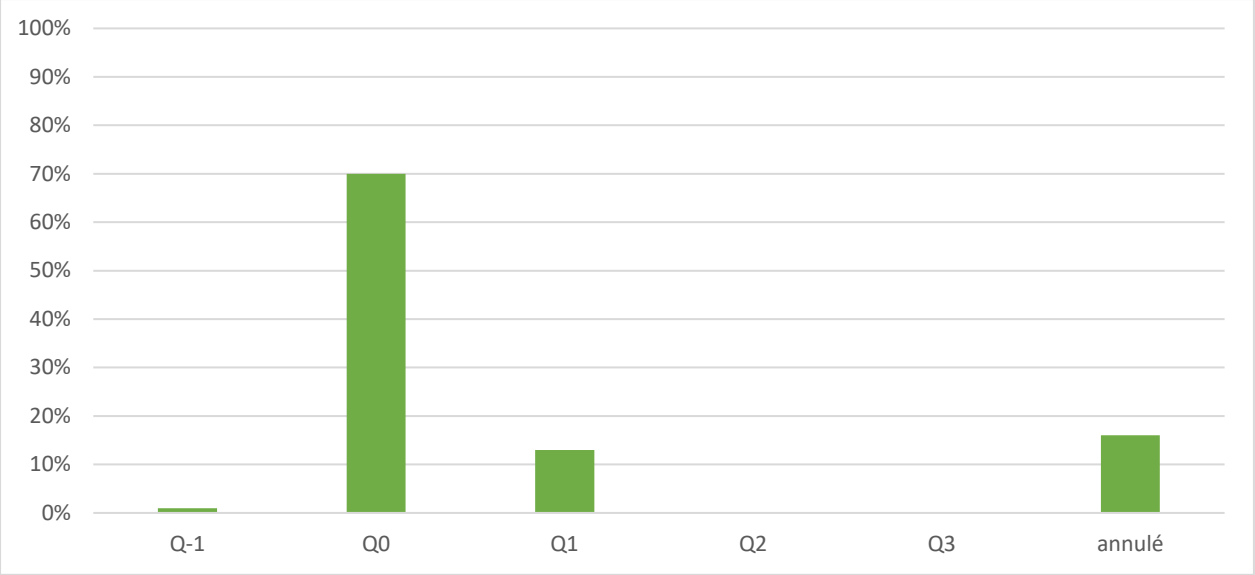

Figure 6.15 : Probabilité prédite pour un modèle des différentes classes pour un devis

A la Figure 6.15, par exemple le devis a une probabilité significativement plus importante d'être vendu en Q0. Il faudra cependant faire attention au risque d'annulation qui s'élève à 15 % selon le modèle. La récolte de ce type d'informations, pour l'ensemble des devis en cours pour un client donné, pourrait permettre de mieux anticiper son annulation ou son retard de réalisation. L'utilisation des modèles détaillés peut aider à la classification des devis en plus de la probabilité donnée par le vendeur.

6.2.2.3.2 Utilisation des modèles simples

En plus des modèles détaillés, les modèles simples peuvent aider à prédire les annulations de devis pour certains clients. En effet, ce type de modèle est plus efficace pour les clients 11 et 12. Il est moins informatif, mais peut permettre d'avoir des résultats efficaces et plus rapides à mettre en place.

Les modèles développés dans cette étude ne permettent pas une classification optimale. Cependant ils pourront être utilisés en complément des indicateurs des vendeurs pour permettre l'amélioration de l'analyse des devis en cours de négociation.

CHAPITRE 7 CONCLUSION ET RECOMMANDATIONS

Cette étude a permis de soulever un problème industriel peu étudié. En effet, les devis sont peu utilisés dans le domaine de l'étude des ventes et de la prévision de la demande. Nous avons pu travailler avec des données uniques de l'entreprise partenaire. Elles étaient complexes et ont nécessité un grand travail de compréhension et de nettoyage pour les clarifier au maximum et permettre des analyses poussées.

L'analyse de la base de données à travers différentes méthodes nous a permis d'aboutir à un premier niveau de conclusions sur le fonctionnement des processus devis chez le partenaire et ses limites. Par la suite, nous avons pu modéliser de plusieurs manières, le comportement des devis pour permettre une prédiction fiable de leurs états finaux. Les résultats obtenus nous permettent désormais d'anticiper si un devis évolue en vente ferme. Toutefois, les choix pris dans notre étude ont pu entraîner des biais et entraîner des limites dans l'analyse des résultats. La réparation du processus d'enregistrement ou la limitation du nombre de clients étudiés sont deux exemples de limitations.

Les modèles présentés permettent cependant d'apporter une aide significative à l'entreprise partenaire qui dispose désormais d'un moyen d'identifier les devis qui présentent un risque d'annulation. Le concept est validé, cependant des études pour généraliser sont à prévoir. Le développement futur de ces modèles permettra à l'entreprise d'améliorer ses prévisions et d'éviter certains risques de surstockage ou de perte de client en cas de non-respect des délais. Ceci pourra améliorer dans les bénéfices de l'entreprise.

Des améliorations de ces modèles sont envisageables. L'obtention d'une meilleure base de données en entrée, la définition de différentes variables de décisions ou d'une échelle de temps plus longues sont autant de pistes d'études pour optimiser les prévisions. Des analyses approfondies des modèles avec de l'« explainable IA » permettraient une meilleure compréhension des modèles basés sur des forêts aléatoires. L'étape suivante serait le développement d'outils construits autour de réseaux de neurones et l'augmentation du nombre de clients dans l'étude.

RÉFÉRENCES

[1] A. Abedi et W. Zhu, «An advanced order acceptance model for hybrid production strategy,» *Journal of Manufacturing Systems,* vol. 55, pp. 82-93, 2020.

[2] H. v. Ooijen et J. Bertrand, «Economic due-date setting in job-shops based on routing and workload dependent flow time distribution functions,» *International Journal of Production Economics,* vol. 74, pp. 261-8, 2001 .

[3] L. Wu, J. Y. Yan et Y. J. Fan, «Data mining algorithms and statistical analysis for sales data forecast,» chez *2012 5th International Joint Conference on Computational Sciences and Optimization, CSO 2012, June 23, 2012 - June 26, 2012*, Harbin, Heilongjiang, China, 2012.

[4] D. Nurbakova et T. Saumet, «Deal Closure Prediction based on User's Browsing Behaviour of Sales Content,» chez *20th IEEE International Conference on Data Mining Workshops, ICDMW 2020, November 17, 2020 - November 20, 2020*, Virtual, Sorrento, Italy, 2020.

[5] R. Legerstee et P. H. Franses, «Do experts' SKU forecasts improve after feedback?,» *Journal of Forecasting,* vol. 33, pp. 69-79, 2014.

[6] P. Chittari et N. R. S. Raghavan, «Support vector based demand forecasting for semiconductor manufacturing,» chez *ISSM 2006 - 15th International Symposium on Semiconductor Manufacturing, September 25, 2007 - September 27, 2007*, Tokyo, Japan, 2006.

[7] A. Pratondo, «Fuzzy rule base for analytical demand forecasting enhancement,» chez *Proceedings - 2010 2nd International Conference on Advances in Computing, Control and Telecommunication Technologies, ACT 2010*, Jakarta, 2010.

[8] W. Fu, C.-F. Chien et Z.-H. Lin, «A hybrid forecasting framework with neural network and time-series method for intermittent demand in semiconductor supply chain,» chez *IFIP WG 5.7 International Conference on Advances in Production Management Systems, APMS 2018, August 26, 2018 - August 30, 2018*, Seoul, Korea, Republic of, 2018.

[9] S.-C. Chang, H.-C. Lai et H.-C. Yu, «A variable P value rolling Grey forecasting model for Taiwan semiconductor industry production,» *Technological Forecasting and Social Change,* vol. 72, pp. 623-640, 2005.

[10] B. Li, J. Li, W. Li et S. A. Shirodkar, «Demand forecasting for production planning decision-making based on the new optimised fuzzy short time-series clustering,» *Production Planning and Control,* vol. 23, pp. 663-673, 2012.

[11] S. Rai, N. Khandelwal et R. Boghey, «Analysis of customer churn prediction in telecom sector using cart algorithm,» chez *1st International Conference on Sustainable Technologies for Computational Intelligence, ICTSCI 2019, March 29, 2019 - March 30, 2019*, Jaipur, India, 2020.

[12] J. Beschi Raja et S. Chenthur Pandian, «An optimal ensemble classification for predicting Churn in

telecommunication,» *Journal of Engineering Science and Technology Review,* vol. 13, pp. 44-49, 2020.

[13] C.-H. Wang et H.-C. Lin, «Competitive substitution and technological diffusion for semiconductor foundry firms,» *Advanced Engineering Informatics,* vol. 48, n° %1101254, 2021.

[14] R.-S. Jiang, E. D. Liou, K.-L. Chen, C.-S. Su et C.-Y. Hung, «Demand forecasting system using customer behavior to optimize foundry manufacturing,» chez *2010 International Symposium on Semiconductor Manufacturing, ISSM 2010*, Tokyo, Japan, 2010.

[15] V. C. Nwaogu et K. Dimililer, «Customer Churn Prediction for Business Intelligence Using Machine Learning,» chez *3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications, HORA 2021, June 11, 2021 - June 13, 2021*, Ankara, Turkey, 2021.

[16] I. Tirkel, «Forecasting flow time in semiconductor manufacturing using knowledge discovery in databases,» *International Journal of Production Research,* vol. 51, pp. 5536-5548, 2013.

[17] C.-C. Lin, Y.-H. Tang, J. Z. Shyu et Y.-M. Li, «A diffusion model to growth phase forecasting of 3G industry in Taiwan,» chez *2008 Portland International Center for Management of Engineering and Technology, Technology Management for a Sustainable Economy, PICMET '08, July 27, 2008 - July 31, 2008*, Cape Town, South africa, 2008.

[18] A. Alamsyah et N. Salma, «A Comparative Study of Employee Churn Prediction Model,» chez *4th International Conference on Science and Technology, ICST 2018, August 7, 2018 - August 8, 2018*, Yogyakarta, Indonesia, 2018.

[19] A. Ahmad, A. Jafar et K. Aljoumaa, «Customer churn prediction in telecom using machine learning in big data platform,» *Journal of Big Data,* vol. 6, p. 28 (24 pp.), 2019.

[20] Y.-J. Chen et C.-F. Chien, «An empirical study of demand forecasting of non-volatile memory for smart production of semiconductor manufacturing,» *International Journal of Production Research,* vol. 56, pp. 4629-4643, 2018.

[21] A. Seitz, M. Grunow et R. Akkerman, «Data driven supply allocation to individual customers considering forecast bias,» *International Journal of Production Economics,* vol. 227, n° %1107683, 2020.

[22] C. Terwiesch, Z. J. Ren, T. H. Ho et M. A. Cohen, «An empirical analysis of forecast sharing in the semiconductor equipment supply chain,» *Management Science,* vol. 51, pp. 208-220, 2005.

[23] K. Knoblich, C. Heavey et P. Williams, «Quantitative analysis of semiconductor supply chain contracts with order flexibility under demand uncertainty: A case study,» *Computers and Industrial Engineering,* vol. 87, pp. 394-406, 2015.

[24] Q. Xu et V. Sharma, «Ensemble Sales Forecasting Study in Semiconductor Industry,» chez *Advances in Data Mining: Applications and Theoretical Aspects. 17th Industrial Conference, ICDM 2017*, New York, NY, USA, 2017.

[25] F. Zhang, «An application of vector GARCH model in semiconductor demand planning,» *European Journal of Operational Research,* vol. 181, pp. 288-297, 2007.

[26] K. C. So et X. Zheng, «Impact of supplier's lead time and forecast demand updating on retailer's order quantity variability in a two-level supply chain,» *International Journal of Production Economics,* vol. 86, pp. 169-179, 2003.

[27] J. M. Framinan et P. Perez-Gonzalez, «Available-To-Promise systems in the semiconductor industry: A review of contributions and a preliminary experiment,» chez *2016 Winter Simulation Conference, WSC 2016, December 11, 2016 - December 14, 2016*, Arlington, VA, United states, 2016.

[28] C.-F. Chien et K.-Y. Lin, «Manufacturing intelligence for Hsinchu Science Park semiconductor sales prediction,» *Journal of the Chinese Institute of Industrial Engineers,* vol. 29, pp. 98-110, 2012.

[29] N. Jain et V. Srivastava, «DATA MINING TECHNIQUES: A SURVEY PAPER,» *International Journal of Research in Engineering and Technology,* vol. 02, pp. 116-119, 2013.

[30] C.-H. Wang et J.-Y. Chen, « Demand forecasting and financial estimation considering the interactive dynamics of semiconductor supply-chain companies,» *Computers & Industrial Engineering,* vol. 138, pp. 132-141 , 2019 .

[31] P.-H. Hsu, C.-H. Wang, J. Z. Shyu et H.-C. Yu, «A Litterman BVAR approach for production forecasting of technology industries,» *Technological Forecasting & Social Change,* vol. 70, p. 67–82, 2002.

[32] C.-F. Chien, Y.-J. Chen et J.-T. Peng, «Manufacturing intelligence for semiconductor demand forecast based on technology diffusion and product life cycle,» *International Journal of Production Economics,* vol. 128, pp. 496-509, 2010.

[33] C. Wohlin, «Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering,» chez *In Proceedings of the 18th international conference on evaluation and assessment in software engineering*, Londres, 2014.

[34] V. K.-S. A. L. J. S. M. Chavent, «ClustGeo: an R package for hierarchical clustering with spatial constraints,» *Computational Statistics,* vol. 33, pp. 1799-1822, 2018.

ANNEXE A CALCUL DES INDICATEURS

Tableau A.1 : Exemple de matrice de confusion simple

Classe estimée / Classe réelle	Positif	Négatif
Positif	VP	FN
Négatif	FP	VN

VP correspond aux Vrai Positifs ;

FN correspond aux Faux Négatifs ;

FP correspond aux Faux Positifs ;

VN correspond aux Vrai Négatifs.

On appellera P (=VP+FN) le nombre de cas réellement positifs dans les données et N (=FP+VN) le nombre de cas négatifs dans les données. PP correspond à la somme de VP et FP.

La précision permet l'évaluation du modèle à donner de faux positifs. On le calcul de la manière suivante :

$$précision = \frac{VP}{PP}$$

La sensibilité donne la désigne la proportion d'éléments ayant reçu un résultat positif à ce test parmi ceux qui ne sont pas réellement positifs (faux négatifs) :

$$Sensibilité = \frac{VP}{P}$$

La spécificité désigne la proportion d'éléments ayant reçu un résultat négatif à ce test parmi ceux qui ne sont pas réellement négatifs :

$$Spécificité = \frac{VN}{VN + FP} = \frac{VN}{N}$$

La justesse est la proportion de bonne prédiction par rapport aux nombres de cas étudiés. Cet indicateur donne une vue d'ensemble :

$$Justesse = \frac{VP + VN}{P + N}$$

Le score F1 est la moyenne harmonique de la précision et de la sensibilité, il permet d'avoir une vue globale de la classification d'une classe :

$$F1_{score} = 2 * \frac{précision * sensibilité}{précision + sensibilité}$$

Ces cinq indicateurs permettent une meilleure compréhension des données. La classe annulation sera regardée plus précisément, car il s'agit de la classe la plus intéressante pour l'entreprise

ANNEXE B DIFFÉRENTS REGROUPEMENTS

Les différents regroupements selon les méthodes sont disponibles en dessous. Ces regroupements sont légèrement différents entre eux et permettent d'obtenir des résultats plus ou moins bons. La méthode de calcul (Manhattan ou Euclidienne) a un réel impact non négligeable dans le regroupement des clients.

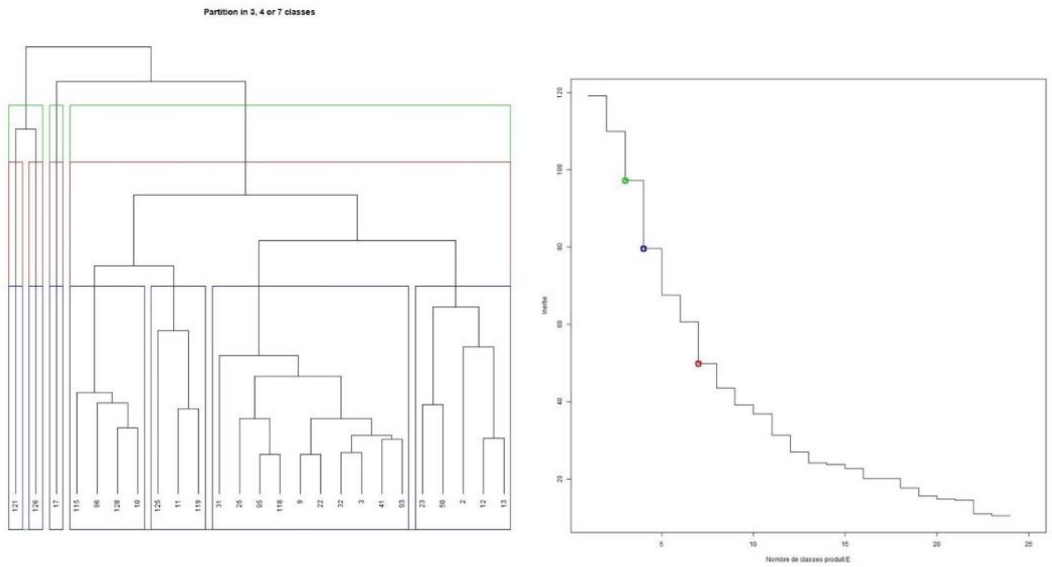

Figure B.1 Trois niveaux de clustering pour la distance produit-euclidienne (gauche) et évolution de l'inertie selon la coupe pour la distance produit-euclidienne (droite)

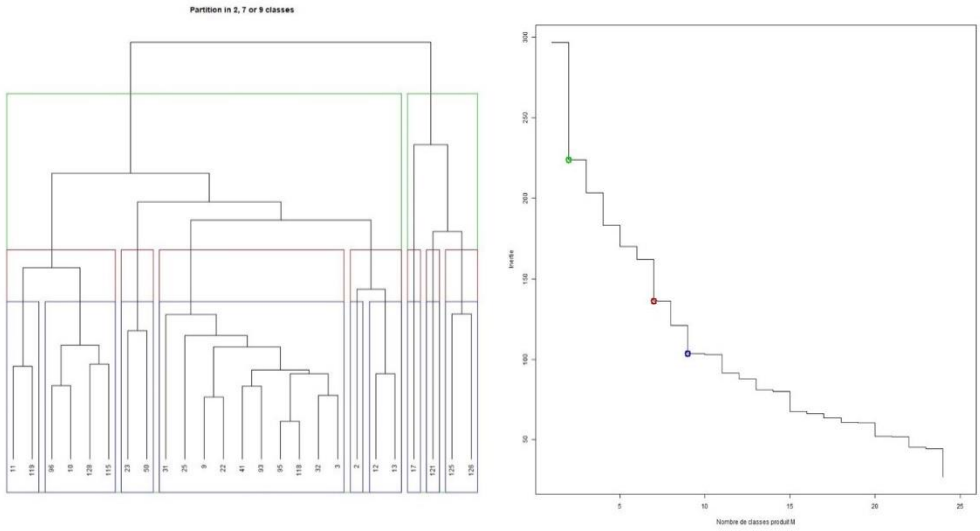

Figure B.2 : Trois niveaux de clustering pour la distance produit-Manhattan (gauche) et évolution de l'inertie selon la coupe pour la distance produit-Manhattan (droite)

ANNEXE C RESULTATS CLIENT 11

1. Justesse globale

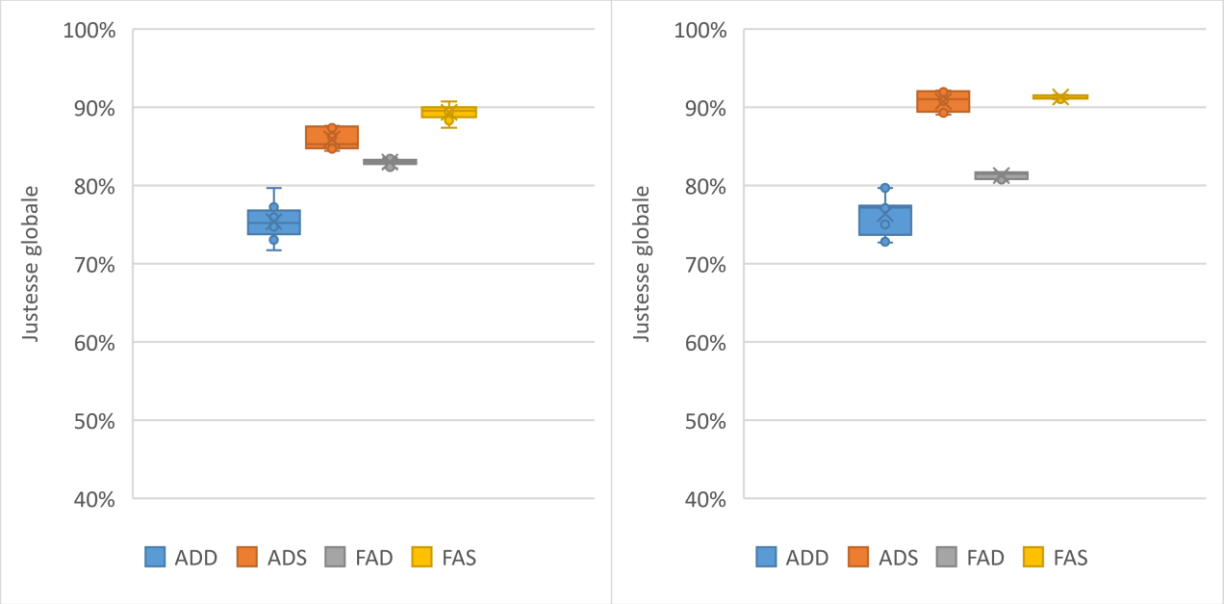

Figure C.1 : Justesse globale (en dollars à gauche et devis à droite) – Client 11

La justesse globale est peu impactée par les regroupements. Les modèles en forêts aléatoires sont meilleurs pour classer les devis, mais les modélisations de l'état final de type simple sont meilleures.

2. Précision, sensibilité, spécificité, F1-score et justesse globale

a. Classe annulation

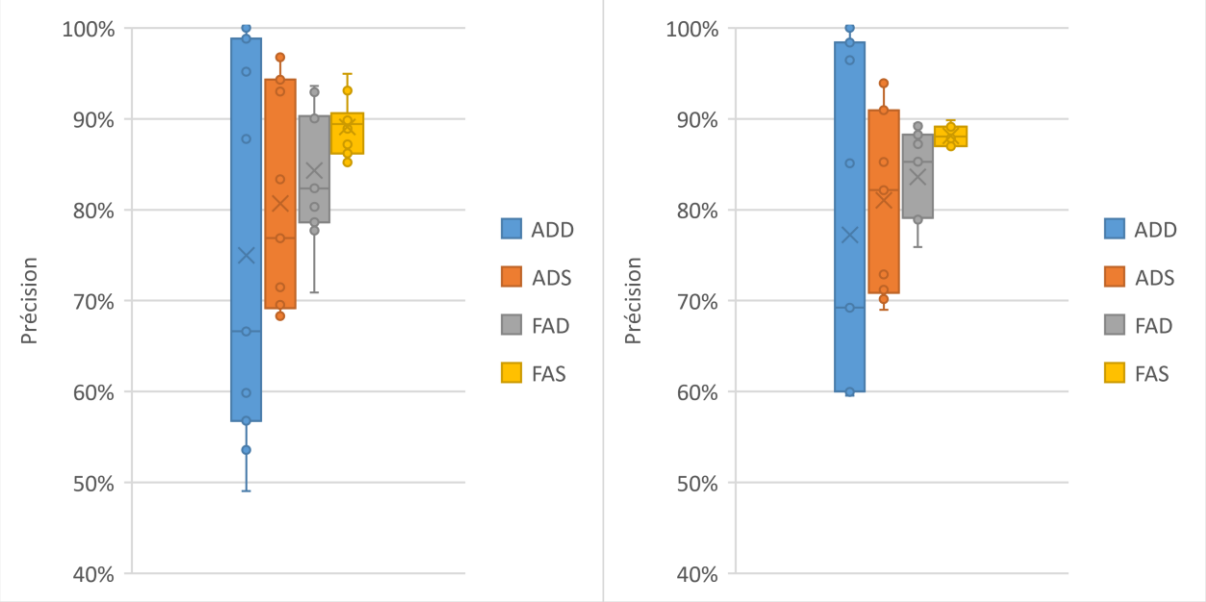

Figure C.2 : Précision en fonction du modèle et du regroupement pour la classe annulation (en dollars à gauche et devis à droite) – Client 11

La précision est fortement impactée par les regroupements. La modélisation FAS est plus stable.

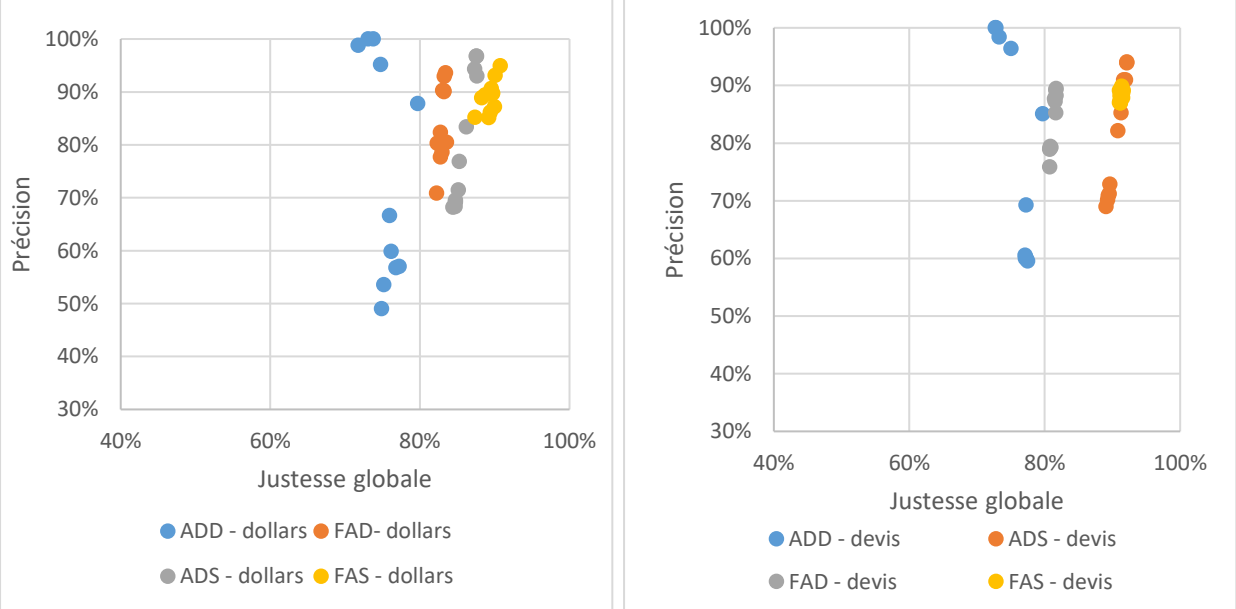

Figure C.3 : Précision de la classe annulation en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 11

La justesse globale est faiblement liée à la précision. L'augmentation de la précision n'est liée qu'à la base d'entraînement.

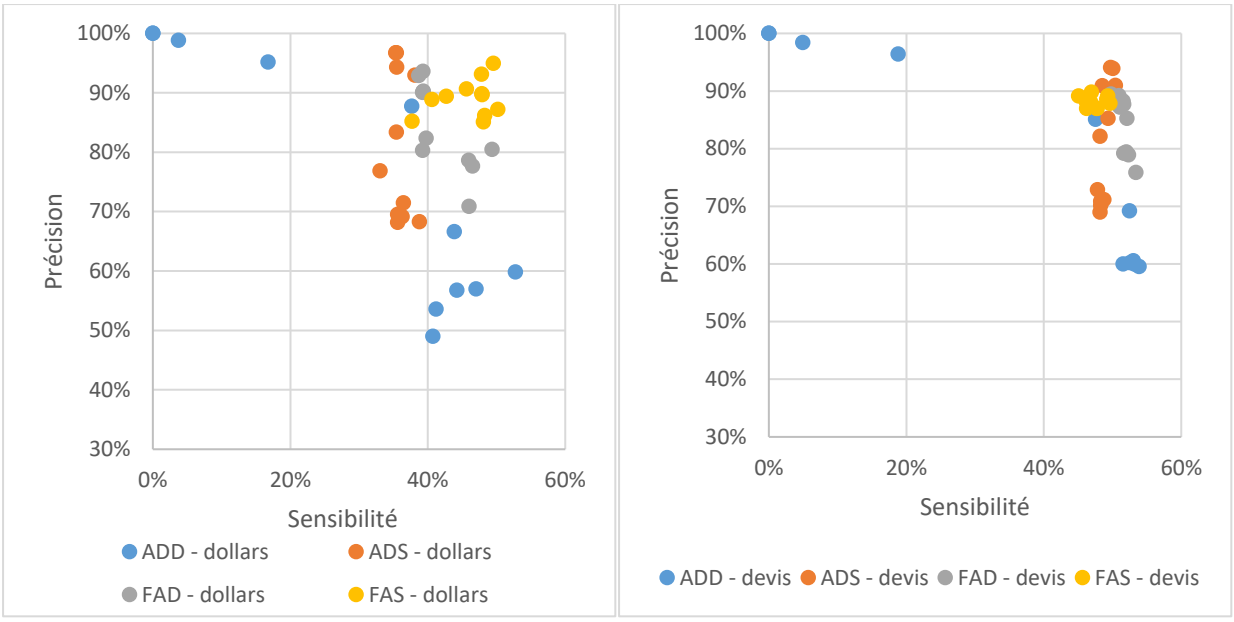

Figure C.4 : Précision de la classe annulation en fonction de la sensibilité (en dollars à gauche et devis à droite) - Client 11

Pour le client 11, les meilleurs modèles arrivent à récupérer 50% des devis annulés avec 95% de précision. (Figure C.4 à gauche). On retrouve bien dans le score F1 (Figure C.5) les bons scores possibles de FAD et de FAS. Le modèle ADD est le modèle le plus fluctuant. Les données ont un fort impact sur l'entraînement et donc sur les prédictions.

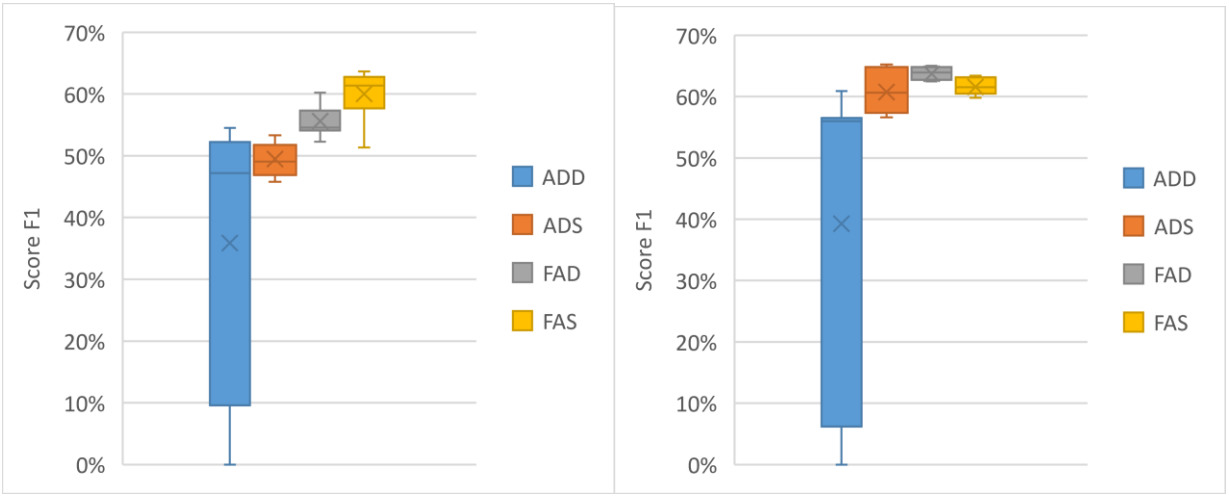

Figure C.5 : Score F1 en fonction du modèle et du regroupement pour la classe annulation (en dollars à gauche et devis à droite) – Client 11

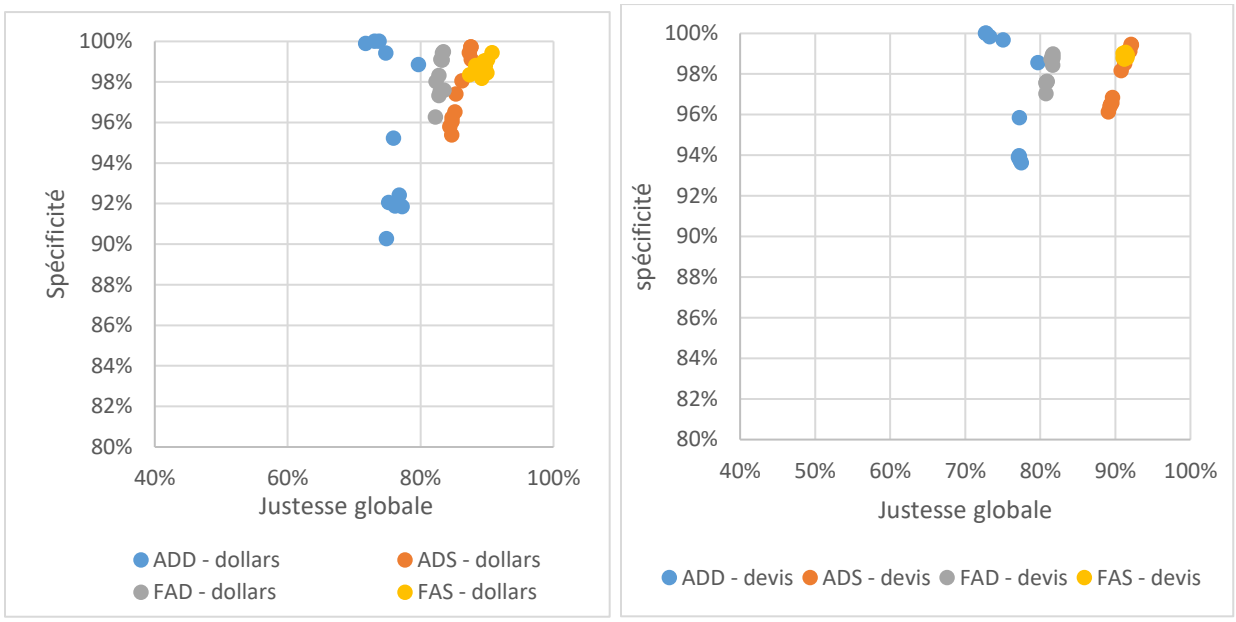

Figure C.6 : Spécificités de la classe annulation en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 11

La spécificité est très bonne, car la majorité des devis sont vendus et bien classés en vendu.

b. Classe vente et vente sur le bon quart

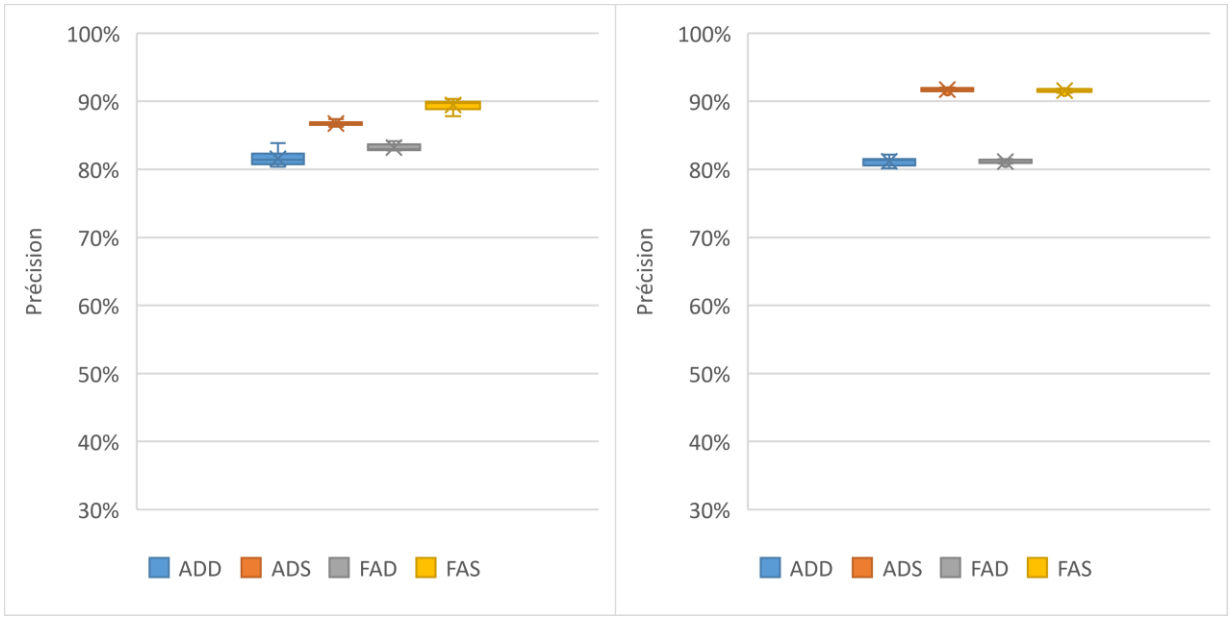

Figure C.7 : Précisions en fonction du modèle et du regroupement pour la classe vente ou Q0 (en dollars à gauche et devis à droite) - Client 11

La précision des modèles est correcte si on compare à la prévalence des devis vendu ou Q0 avec un avantage pour les modèles simples avec un très faible impact des regroupements.

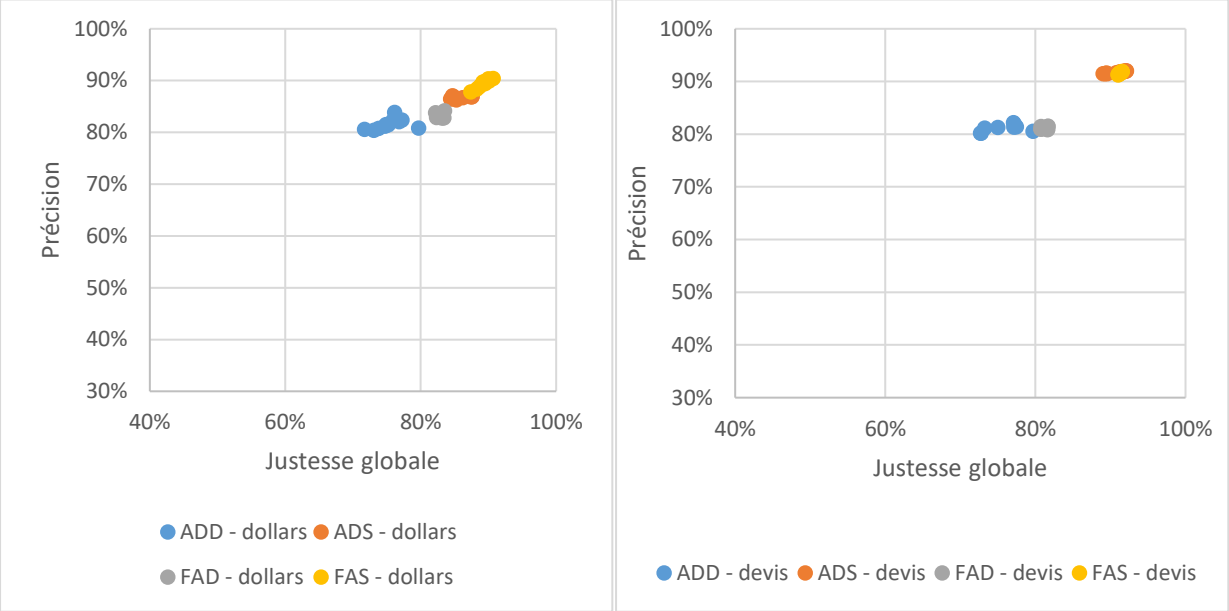

Figure C.8 : Précision de la classe « VENDU» ou Q0 en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 11

La justesse globale est corrélée à la précision. L'augmentation de l'une améliore l'autre.

Pour le client 11, tous les modèles sont équivalents en termes de sensibilité et de précision pour la classe « VENDU» ou Q0.

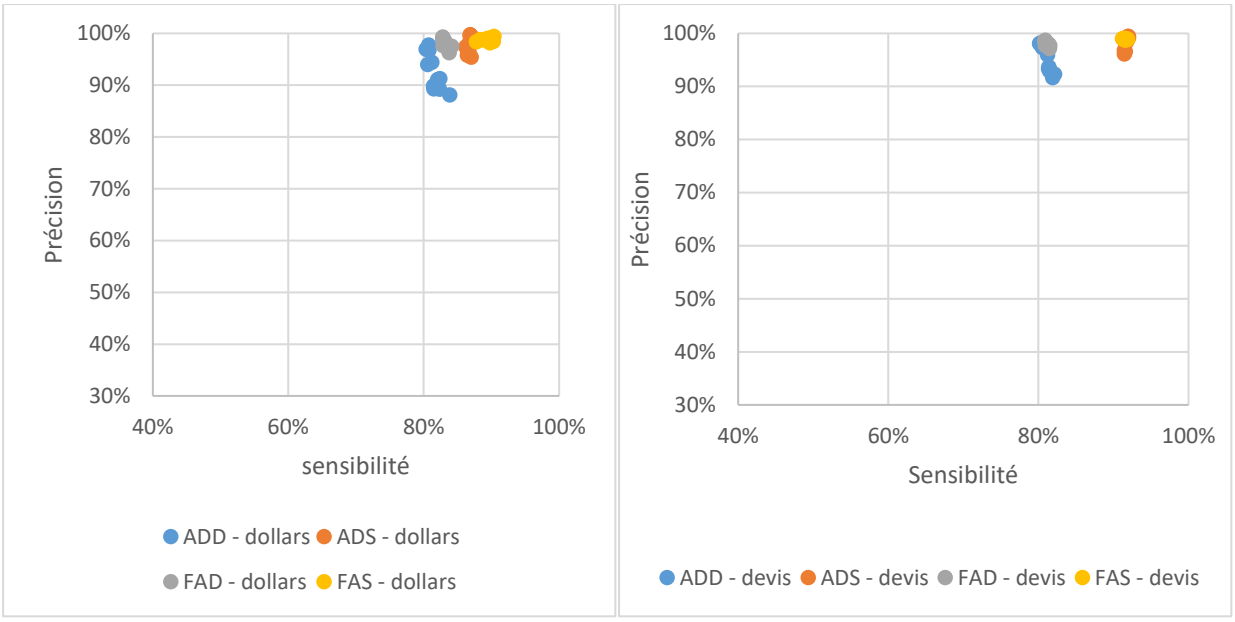

Figure C.9 : Précision de la classe « VENDU » ou Q0 en fonction de la sensibilité (en dollars à gauche et devis à droite) - Client 11

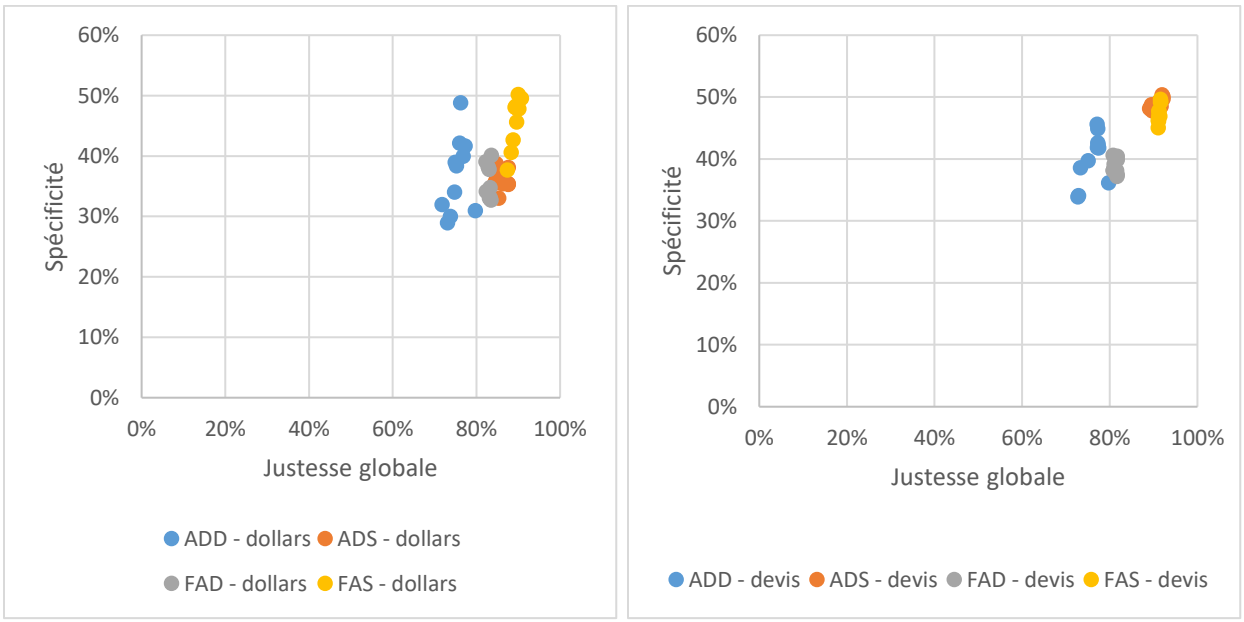

Figure C.10 : Spécificité de la classe « VENDU » ou « Q0 » en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 11

La spécificité de la classe Q0 ou « VENDU » est assez faible, car beaucoup de devis sont encore prédits en Q0 alors qu'ils n'appartiennent pas à cette classe.

ANNEXE D RESULTATS CLIENT 12

1. Justesse globale

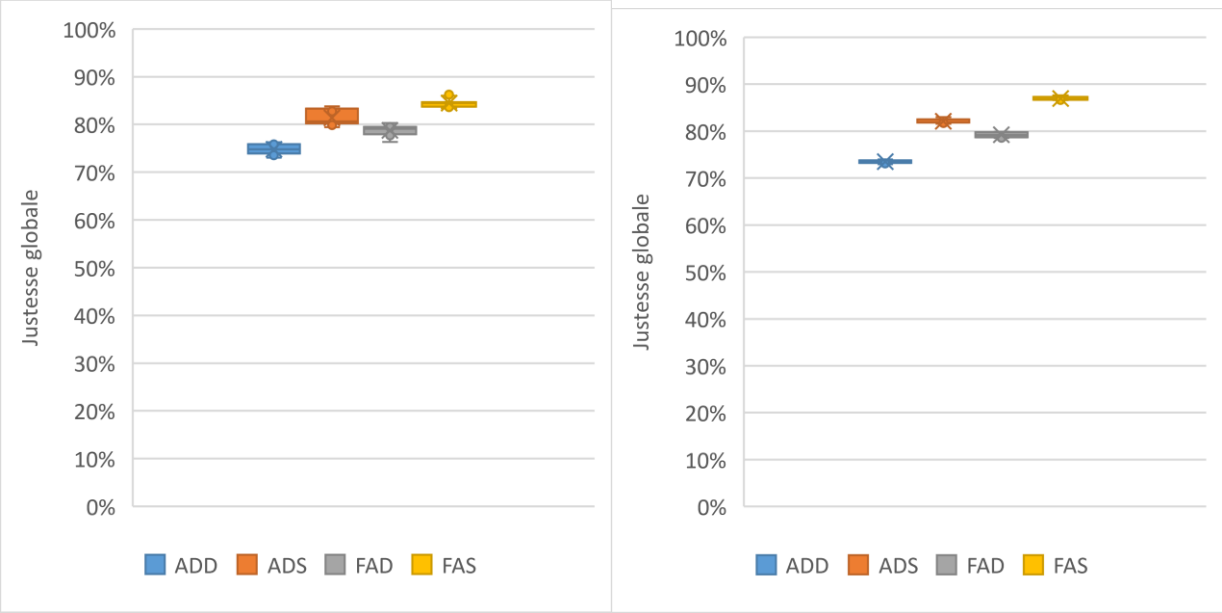

Figure D.1 : Justesse globale (en dollars à gauche et devis à droite) – Client 12

La justesse globale est peu impactée par les regroupements. Les modèles sont très similaires en termes de résultats avec un léger avantage pour les classifications simples.

2. Précision, sensibilité, spécificité, F1-score et justesse globale

a. Classe annulation

La précision est faiblement impactée par les regroupements. On peut lire une variation entre les regroupements d'environ 10% maximum pour le modèle ADS (Figure D.2).

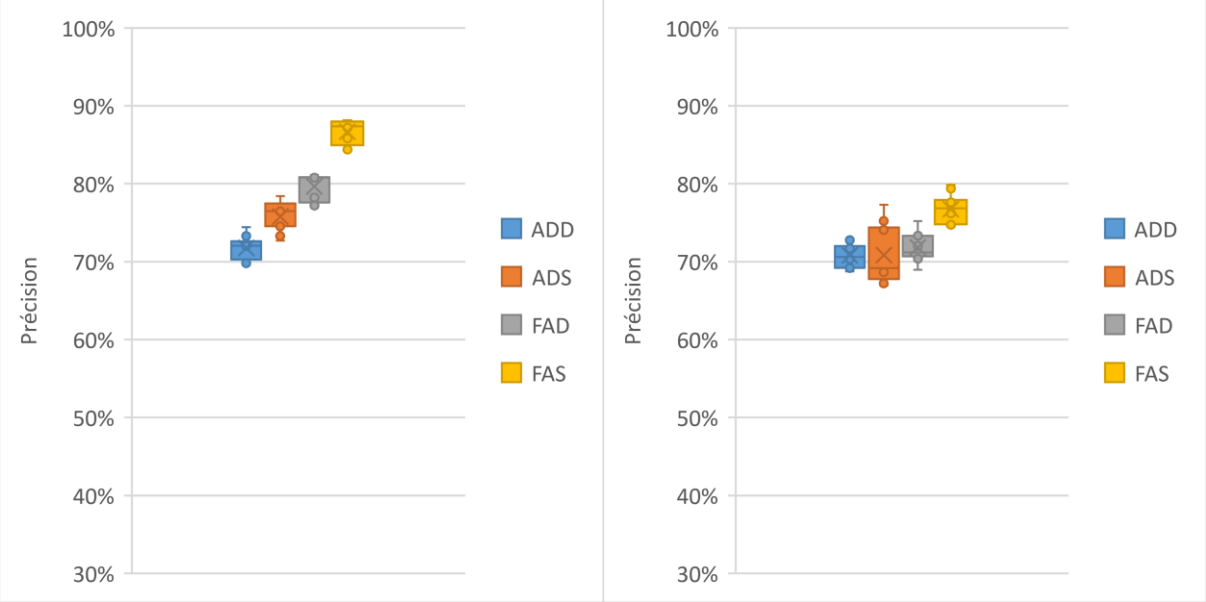

Figure D.2 : Précision en fonction du modèle et du regroupement pour la classe annulation (en dollars à gauche et devis à droite) – Client 12

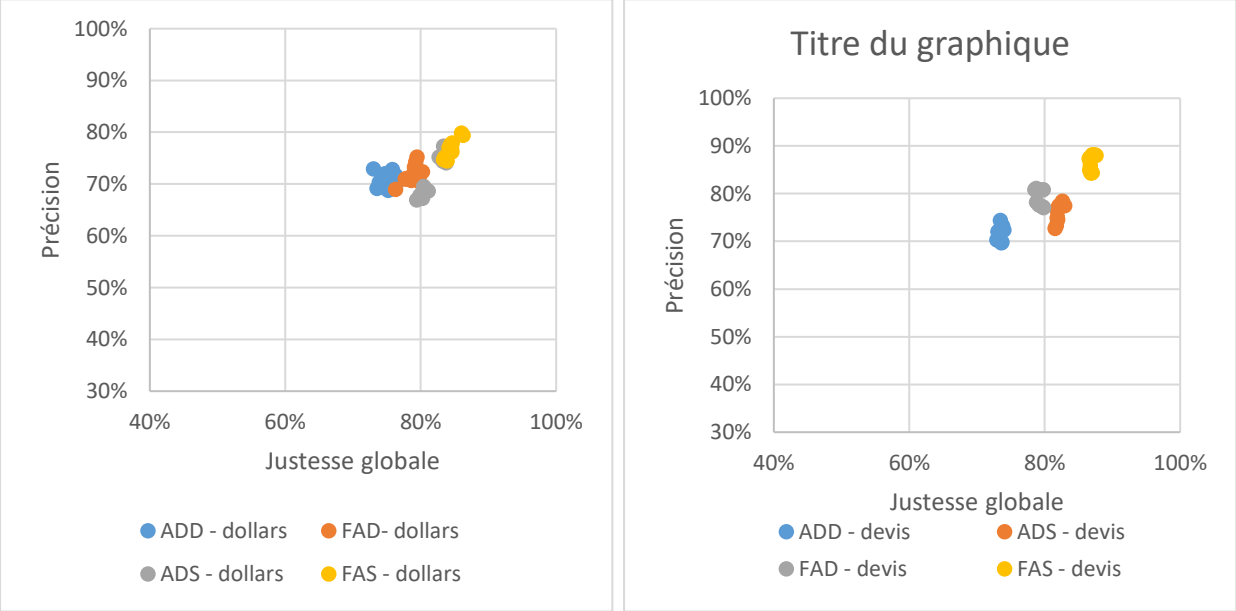

Figure D.3 : Précision de la classe annulation en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 12

La justesse globale est assez indépendante de la précision. L'augmentation de l'une améliore que très légèrement l'autre. Il y a peu de différence entre les résultats. Les regroupements n'impactent que légèrement les résultats de justesse et de précision pour le client 12.

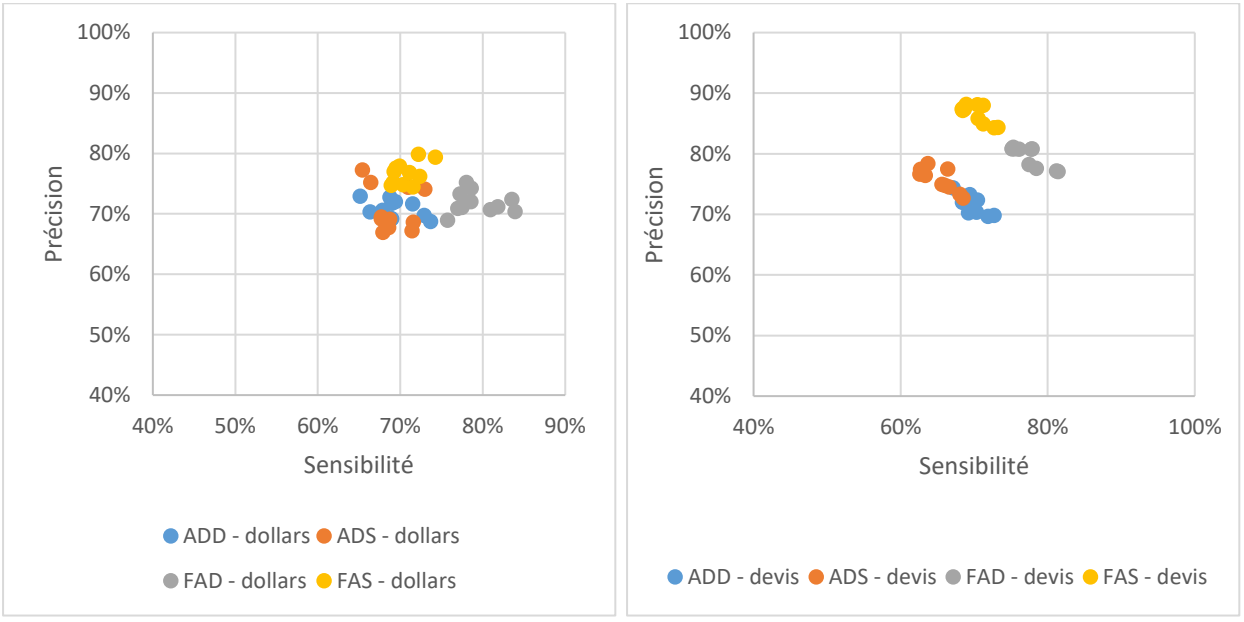

Figure D.4 : Précision de la classe annulation en fonction de la sensibilité (en dollars à gauche et devis à droite) - Client 12

Pour le client 12, les meilleurs modèles arrivent à récupérer 79% des devis annulés avec 74% de précision. (Figure D.4). On retrouve bien dans le score F1 (Figure D.5) les bons scores possibles des modèles FAS et FAD.

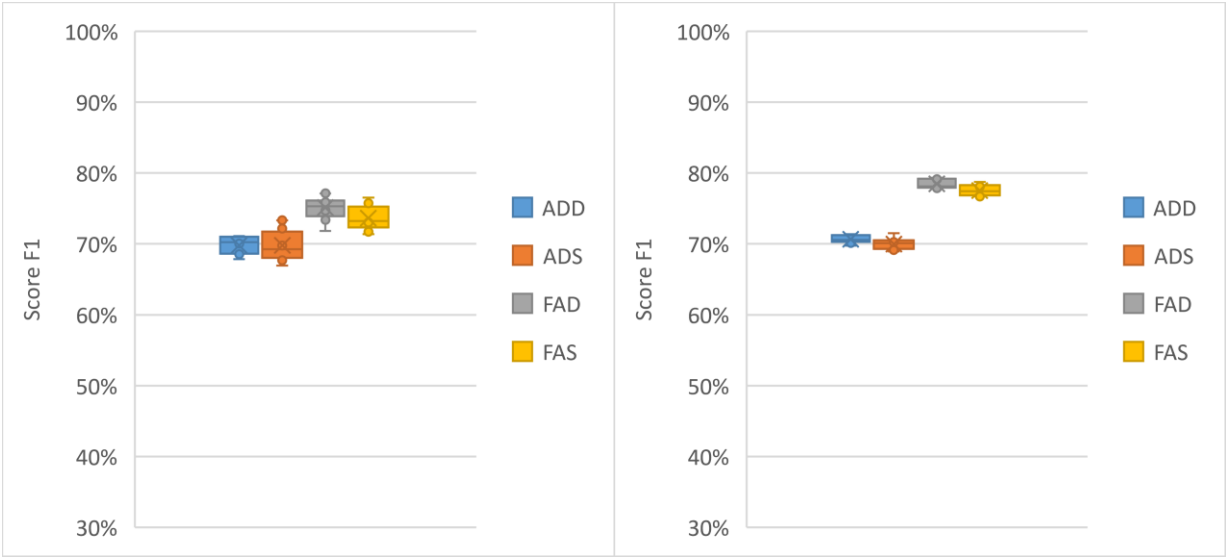

Figure D.5 : Score F1 en fonction du modèle et du regroupement pour la classe annulation (en dollars à gauche et devis à droite) – Client 12

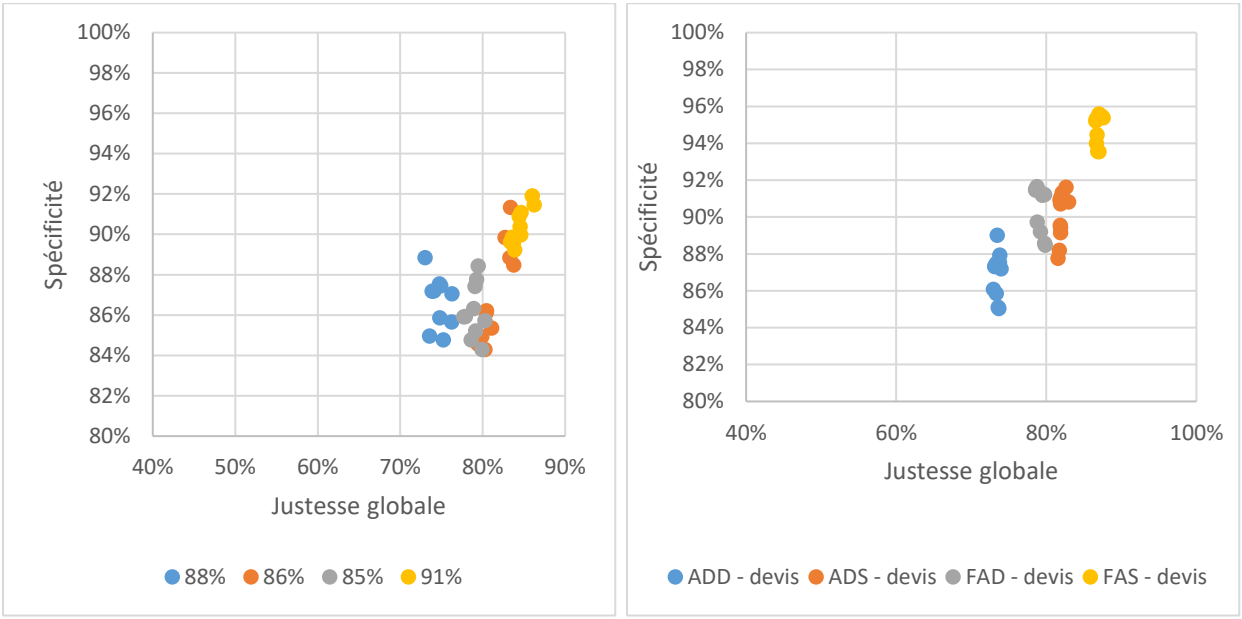

Figure D.6 : Spécificités de la classe annulation en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 12

La spécificité est très bonne, car la majorité des devis sont vendus et bien classés en vendu.

b. Classe vente et vente sur le bon quart

Figure D.7 : Précisions en fonction du modèle et du regroupement pour la classe « VENDU » ou Q0 (en dollars à gauche et devis à droite) - Client 12

La précision des modèles est bonne si on compare à la prévalence des devis « VENDU » ou Q0 avec un avantage pour les modèles simples avec un très faible impact des regroupements.

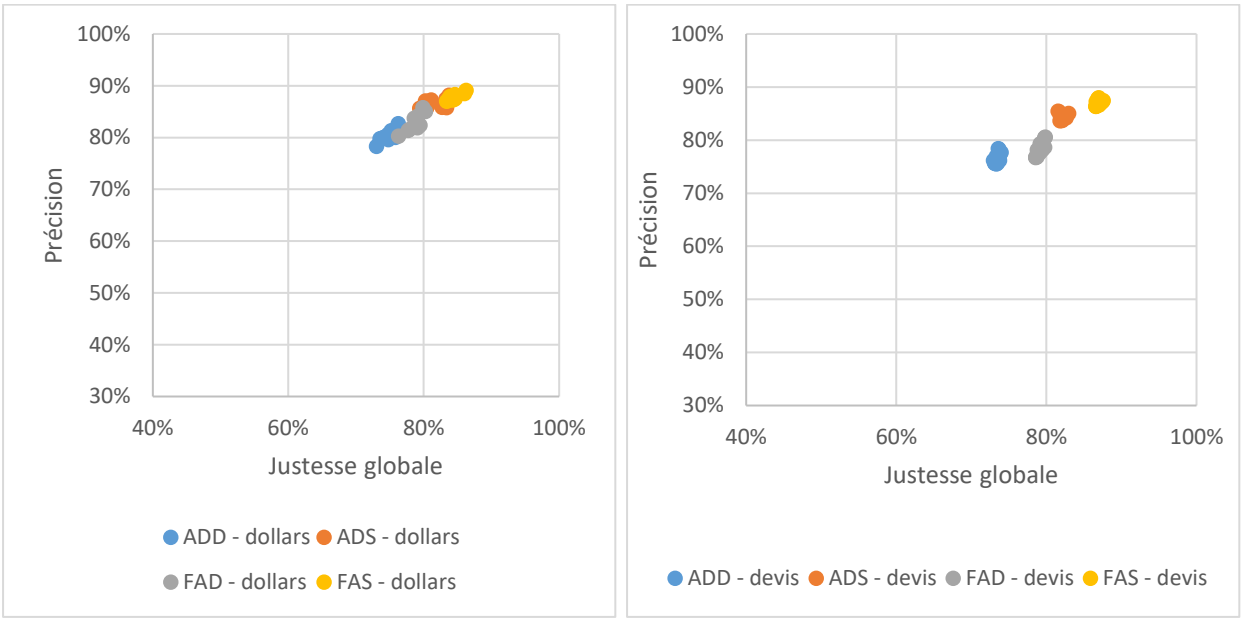

Figure D.8 : Précision de la classe « VENDU » ou Q0 en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 12

Pour le client 12, tous les modèles sont équivalents en termes de sensibilité et de précision pour la classe « VENDU » ou Q0.

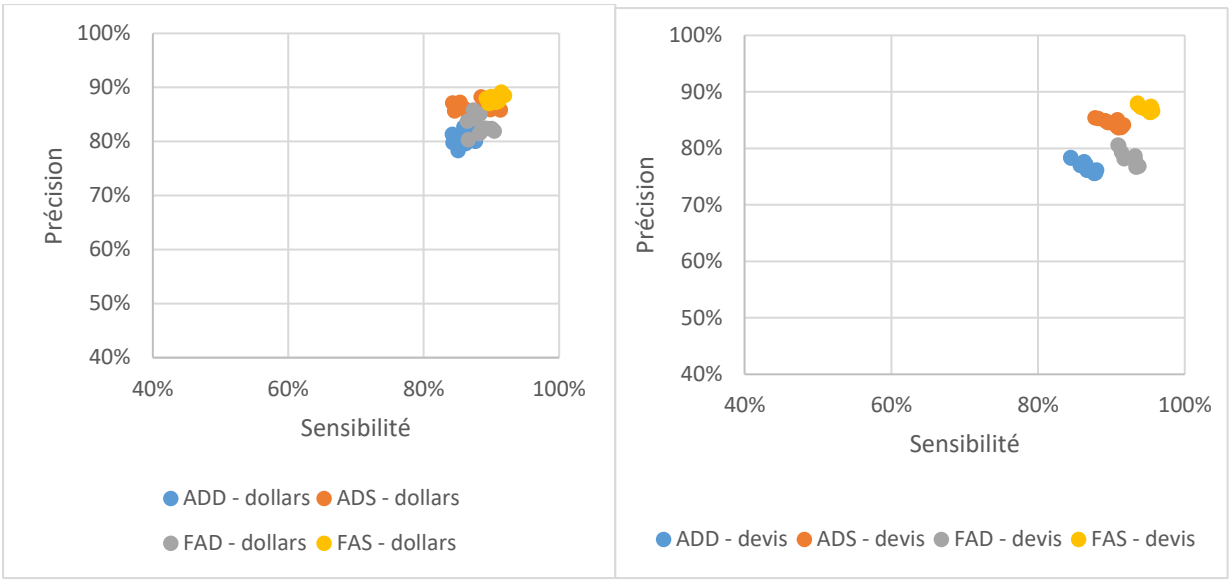

Figure D.9 : Précision de la classe « VENDU » ou Q0 en fonction de la sensibilité (en dollars à gauche et devis à droite) - Client 12

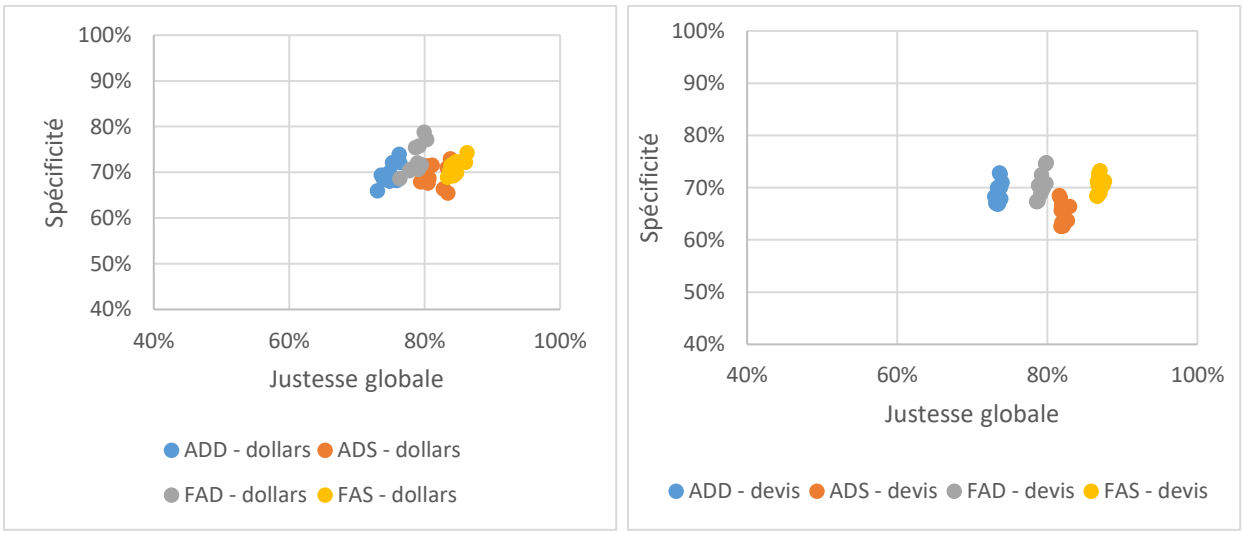

Figure D.10 : Spécificité de la classe « VENDU » ou « Q0 » en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 12

La spécificité de la classe Q0 ou « VENDU » est moyenne, car beaucoup de devis sont encore prédits en Q0 alors qu'ils n'appartiennent pas à cette classe.

ANNEXE E RESULTATS CLIENT 22

1. Justesse globale

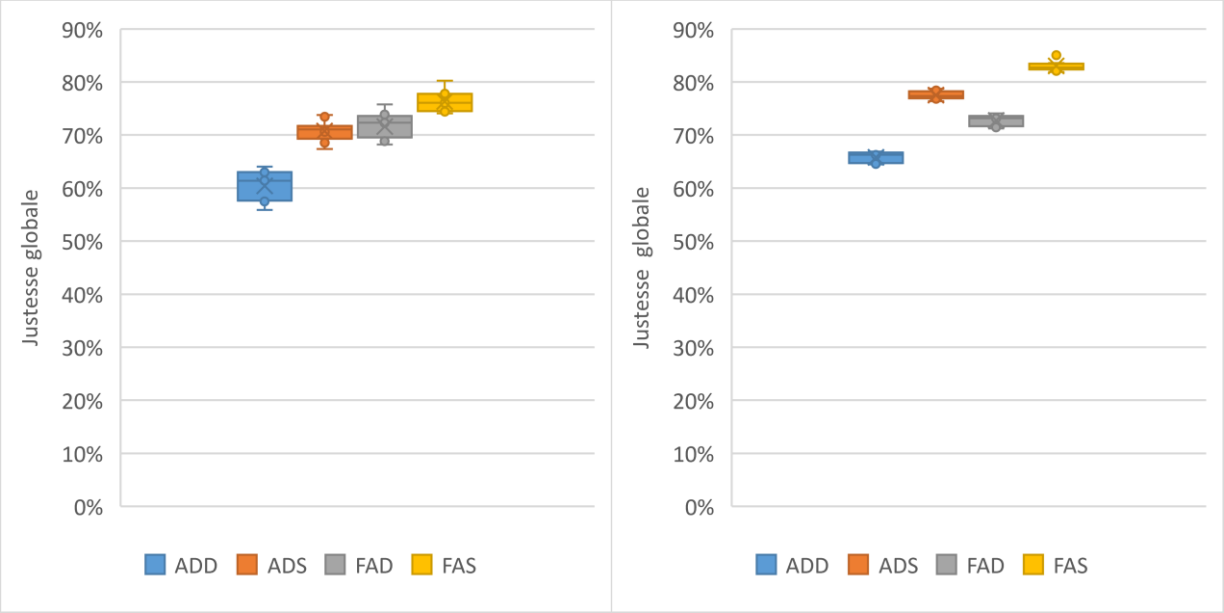

Figure E.1 : Justesse globale (en dollars à gauche et devis à droite) – Client 22

La justesse globale est peu impactée par les regroupements. Les modèles en forêts aléatoires sont meilleurs dans leur catégorie. On peut cependant noter que l'impact du regroupement est plus fort lorsque l'on considère le poids du devis.

2. Précision, sensibilité, spécificité, F1-score et justesse globale

a. Classe annulation

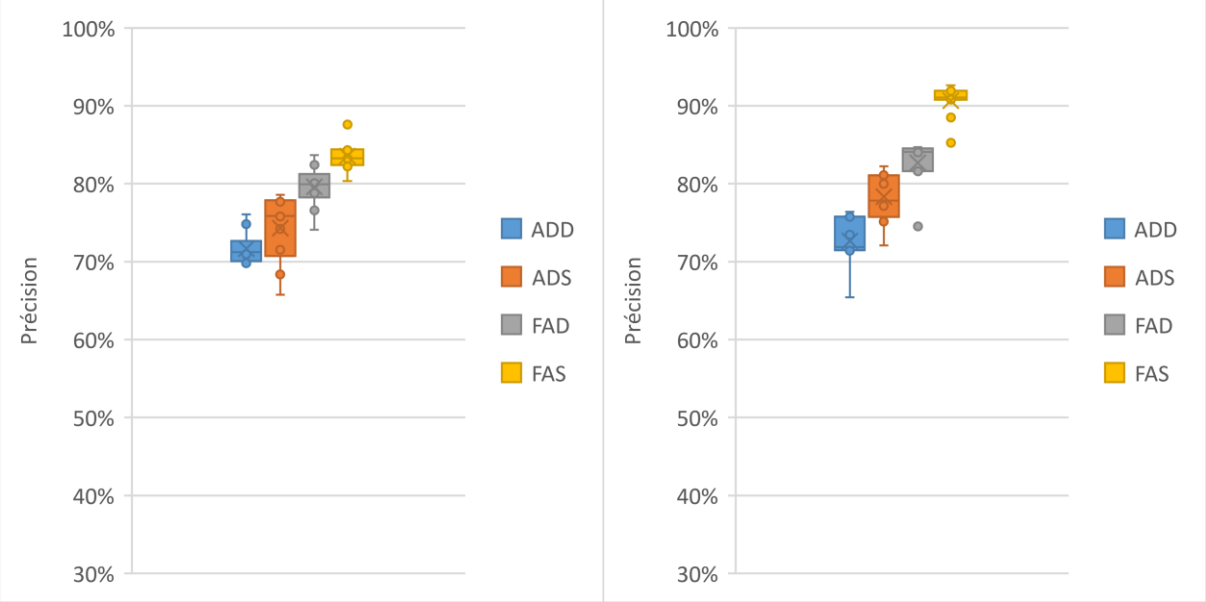

Figure E.2 : Précision en fonction du modèle et du regroupement pour la classe annulation (en dollars à gauche et devis à droite) – Client 22

La précision des modèles « arbre de décision » est fortement impactée par les regroupements. Les modèles en forêts aléatoires sont plus stables et donne les meilleurs résultats. (Figure E.2)

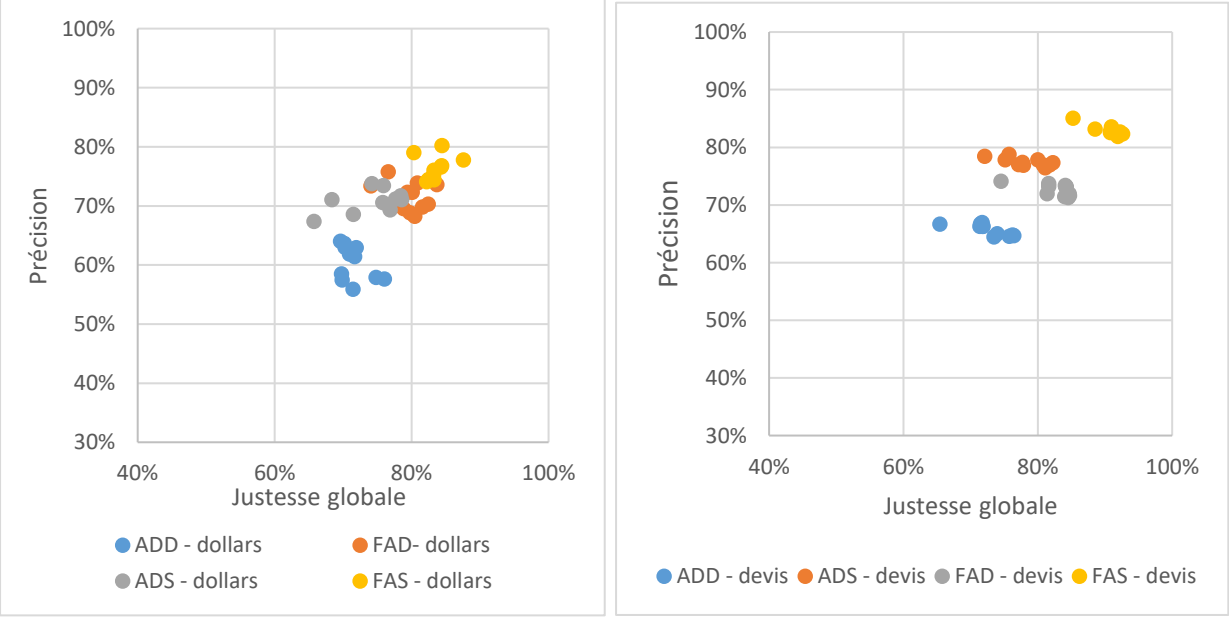

Figure E.3 : Précision de la classe annulation en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 22

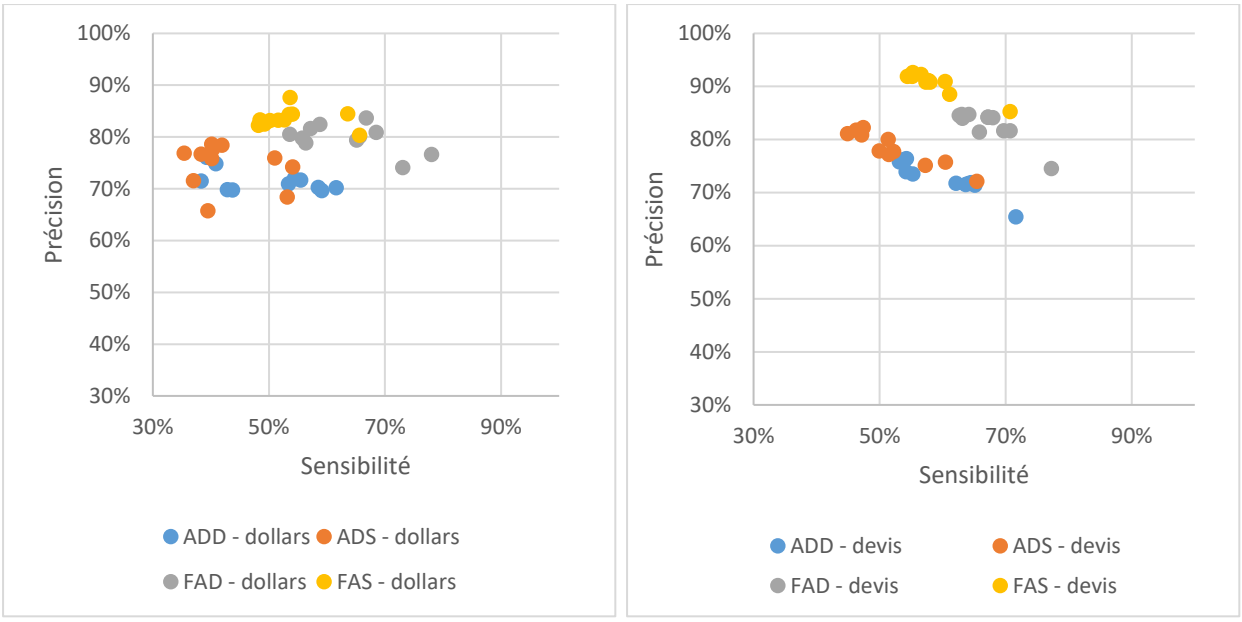

Figure E.4 : Précision de la classe annulation en fonction de la sensibilité (en dollars à gauche et devis à droite) - Client 22

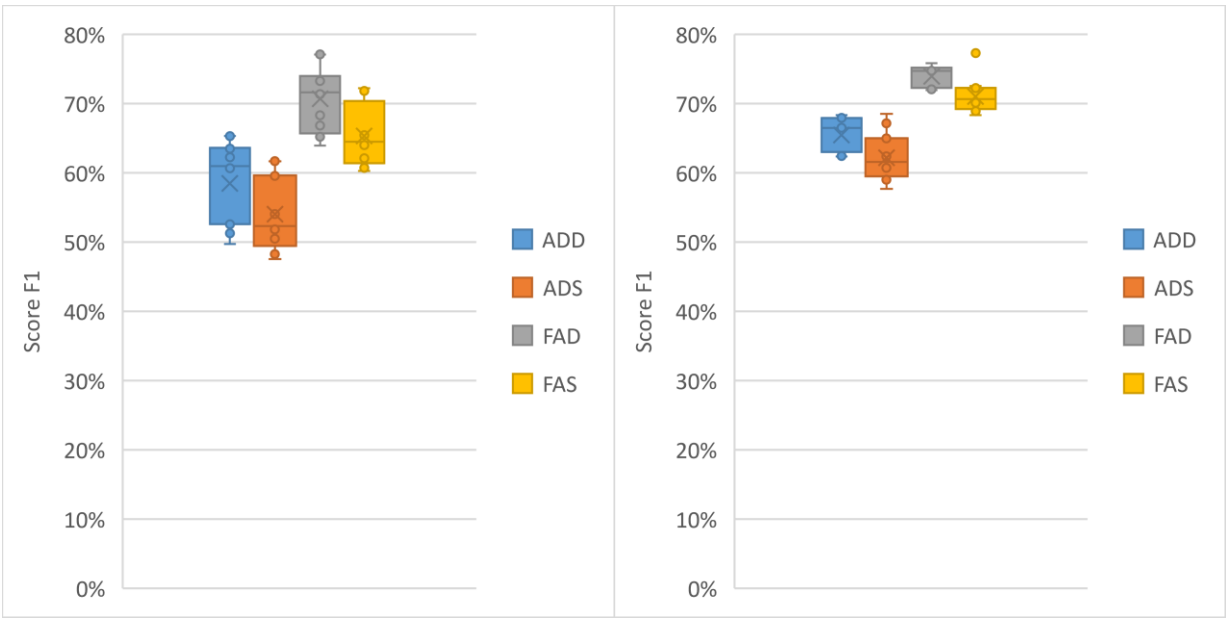

Figure E.5 : Score F1 en fonction du modèle et du regroupement pour la classe annulation (en dollars à gauche et devis à droite) – Client 22

Pour le client 22, les meilleurs modèles arrivent à récupérer 64% des devis annulés avec 84% de précision. (Figure E.4). On retrouve bien dans le score F1 (Figure E.5) les bons scores possibles

du FAD et du FAS. Le regroupement impacte fortement les résultats lorsque l'on considère les poids des devis.

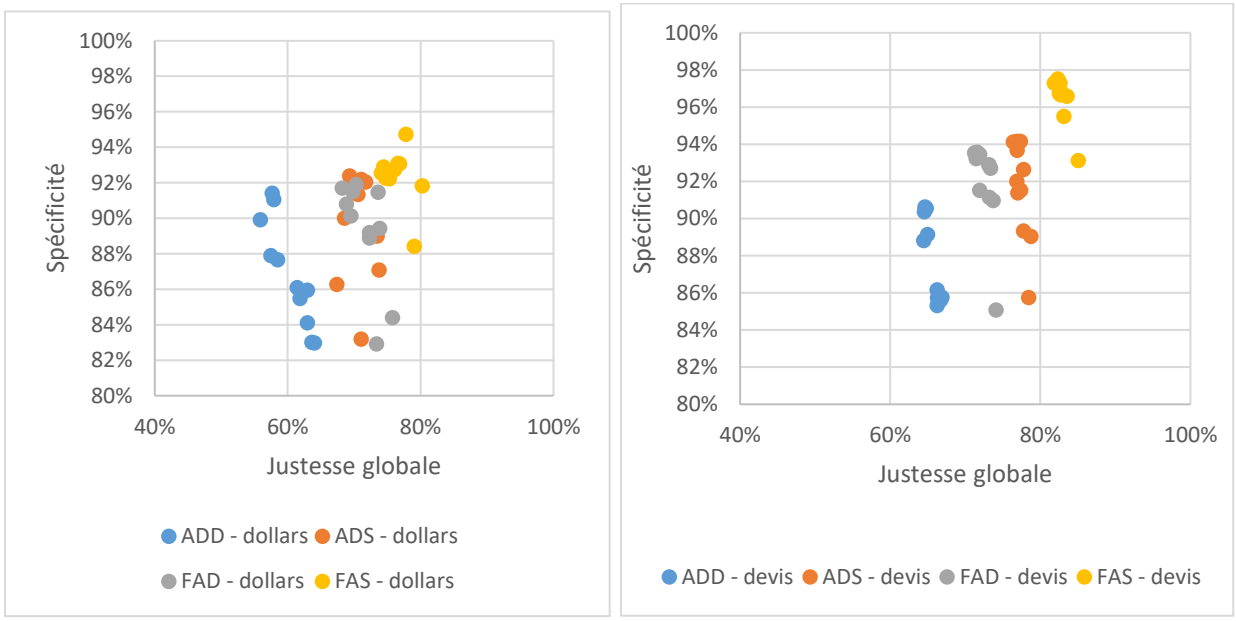

Figure E.6 : Spécificités de la classe annulation en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 22

La spécificité est très bonne malgré une part de devis annulés forte pour ce client. La classification des devis annulés est très bonne pour ce client.

b. Classe vente et vente sur le bon quart

Figure E.7 : Précisions en fonction du modèle et du regroupement pour la classe vente ou Q0 (en dollars à gauche et devis à droite) - Client 22

La précision des modèles est faible et est faiblement impactée par les regroupements. (Figure E.7) Les modèles simples permettent d'obtenir des résultats acceptables. (Figure E.8)

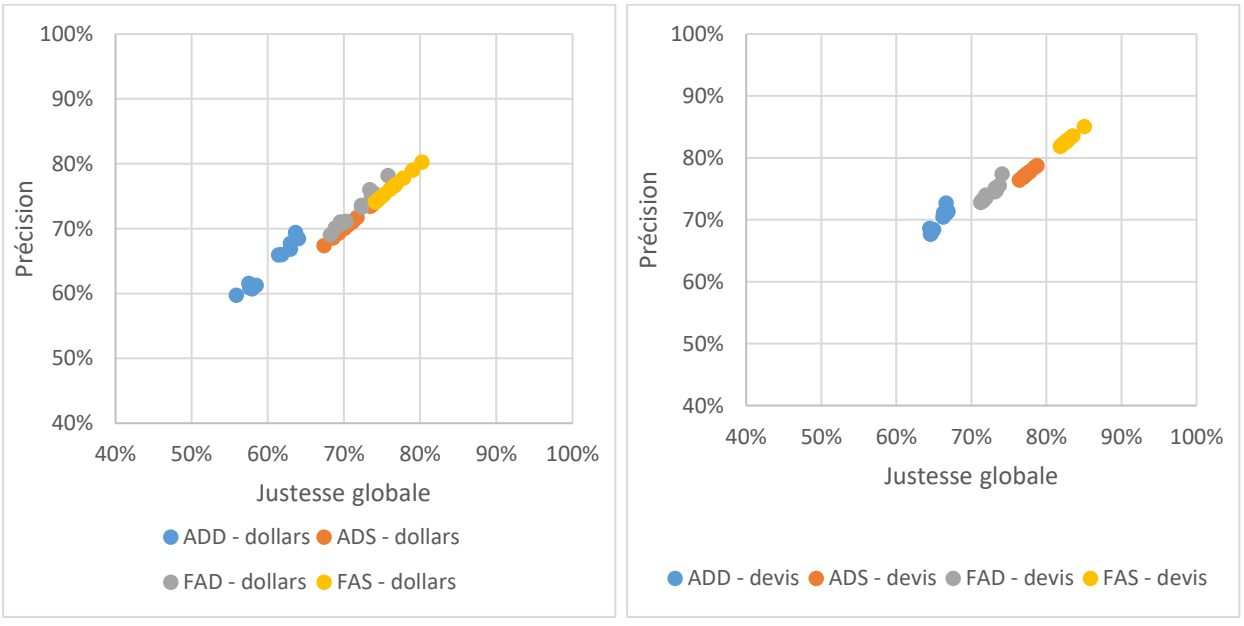

Figure E.8 : Précision de la classe « VENDU » ou Q0 en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 22

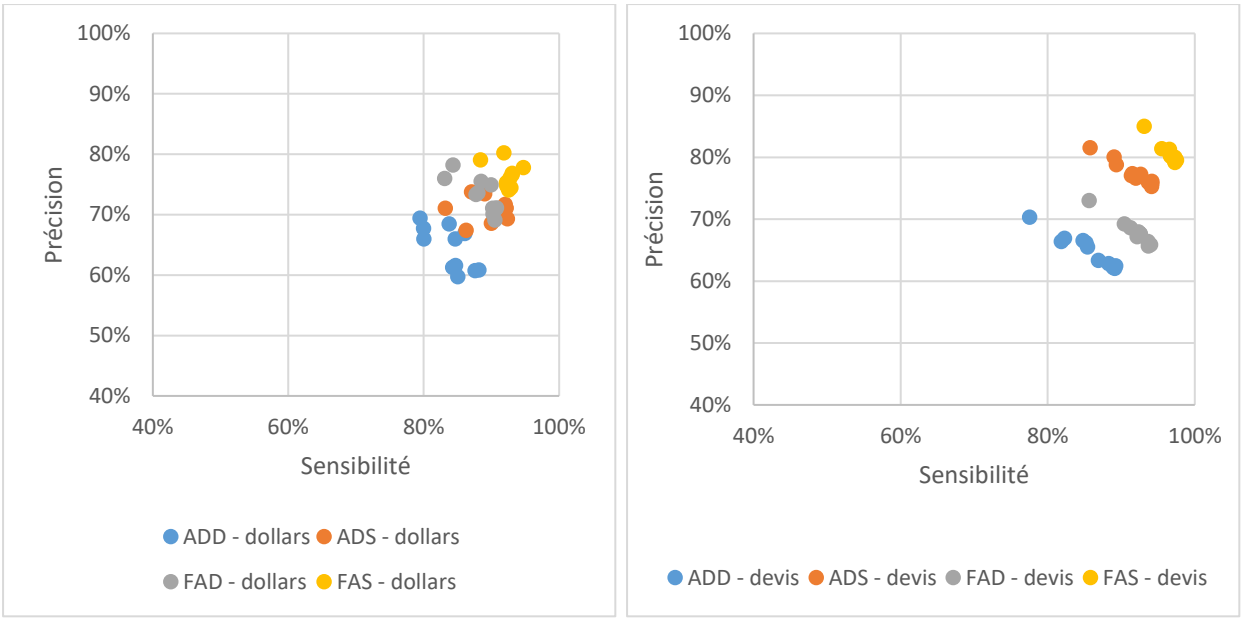

Figure E.9 : Précision de la classe « VENDU » ou Q0 en fonction de la sensibilité (en dollars à gauche et devis à droite) - Client 22

Pour le client 22, tous les modèles simples sont au-dessus en termes de sensibilité et de précision pour la classe « VENDU » ou Q0.

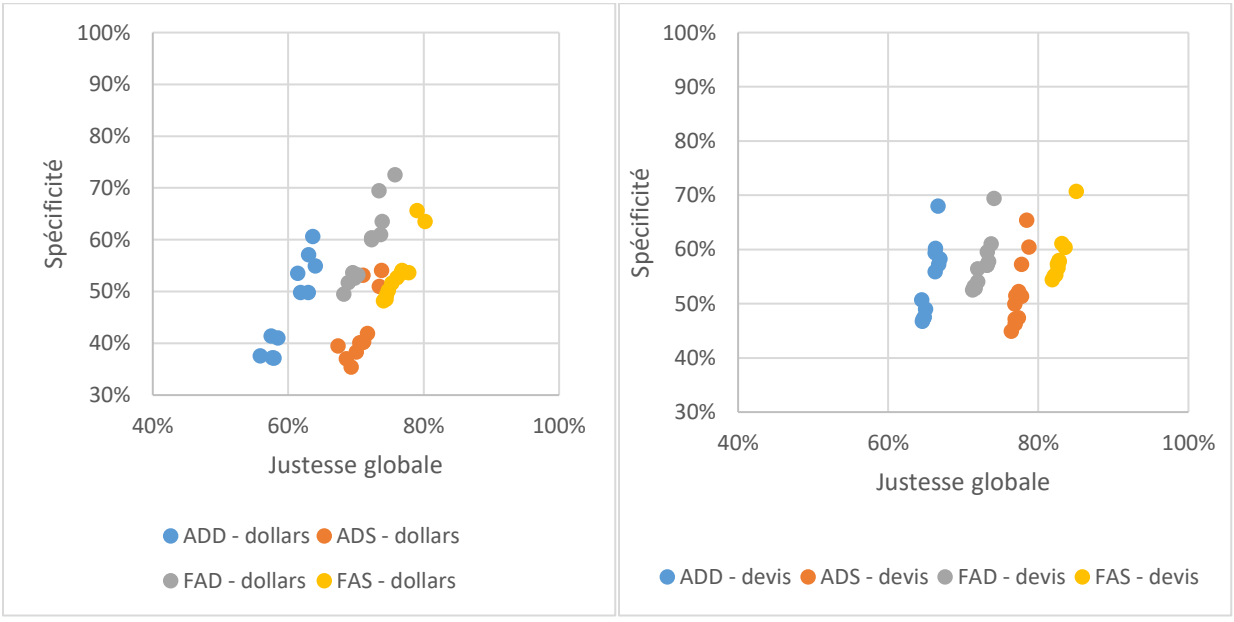

Figure E.10 : Spécificité de la classe « VENDU » ou « Q0 » en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 22

Les différents modèles sont équivalents et permettent d'obtenir une spécificité maximale similaire. Cependant en prenant en compte la justesse en même temps, FAS domine les autres. (Figure E.10 à droite) Si on prend en compte le poids en dollars, FAD passe est meilleur que FAS.

ANNEXE F RESULTATS CLIENT 23

1. Justesse globale

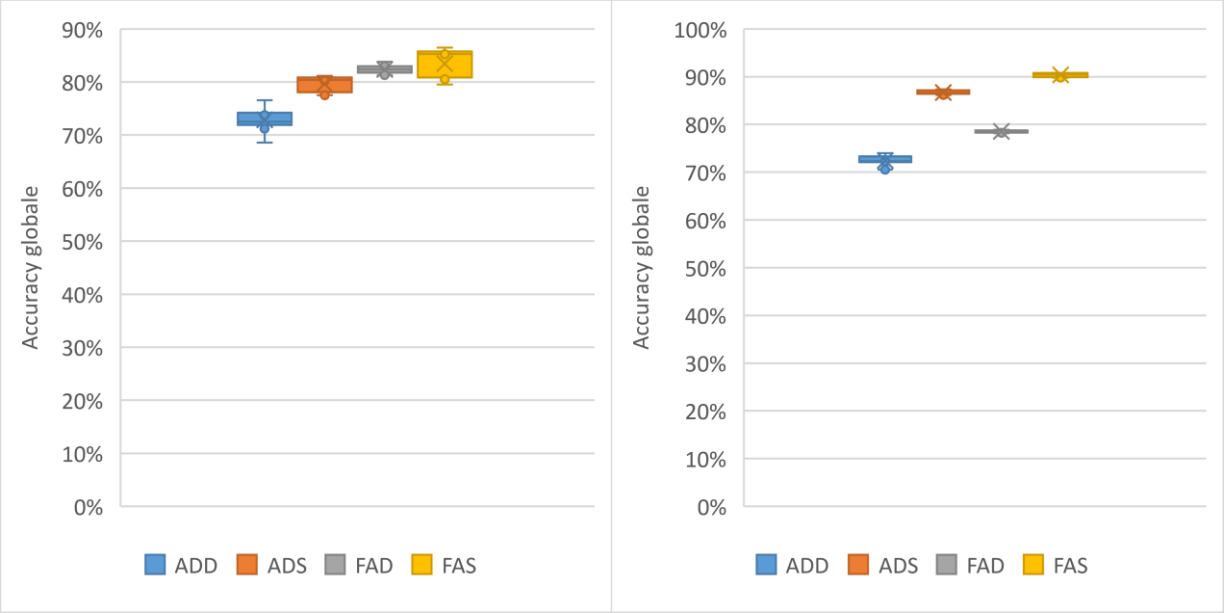

Figure F.1 : Justesse globale (en dollars à gauche et devis à droite) – Client 23

La justesse globale est peu impactée par les regroupements. Les modèles FAD et FAS sont meilleurs au niveau des dollars (Figure F.1). En termes de devis, les modèles simples sont similaires avec peu de variations selon les regroupements.

2. Précision, sensibilité, spécificité, F1-score et justesse globale

a. Classe annulation

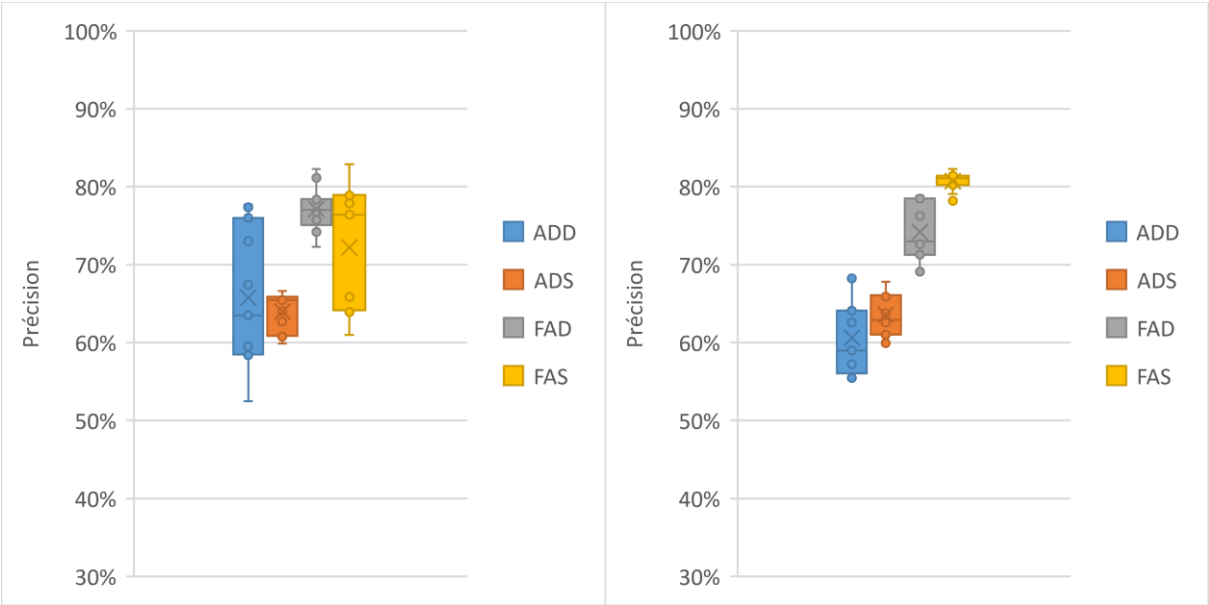

Figure F.2 : Précision en fonction du modèle et du regroupement pour la classe annulation (en dollars à gauche et devis à droite) – Client 23

La précision est fortement impactée par les regroupements pour les modèles ADD et FAS. Les modèles en forêts aléatoires détaillés sont les plus stables au niveau des dollars et permettent d'obtenir de très bons résultats.

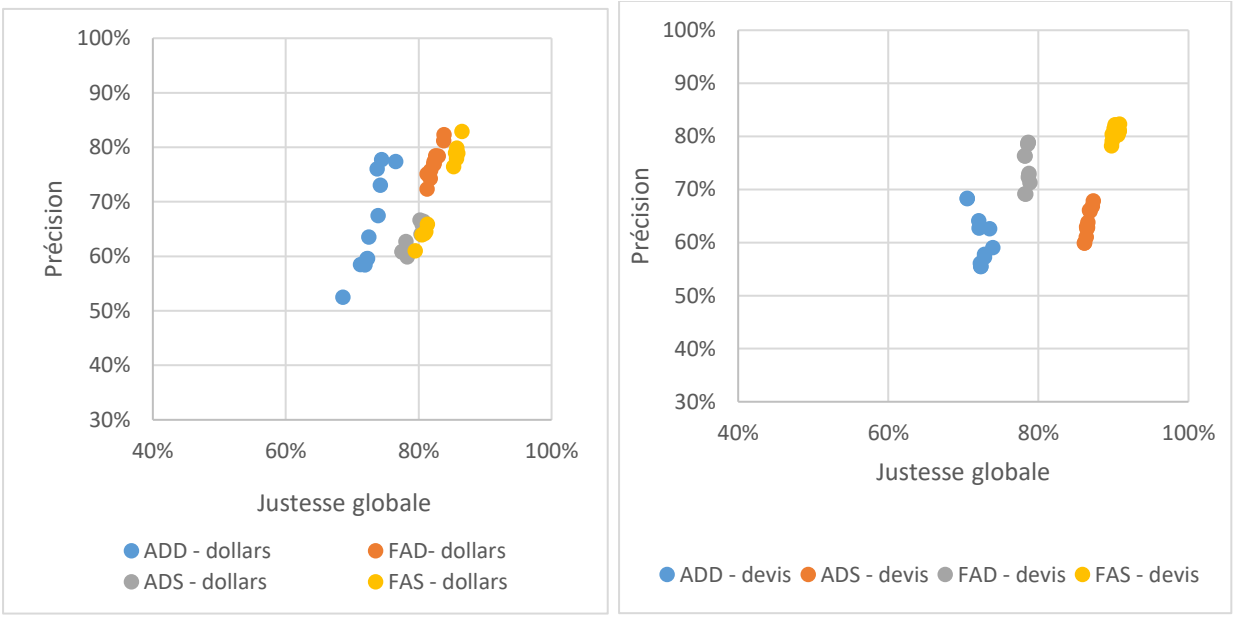

Figure F.3 : Précision de la classe annulation en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 23

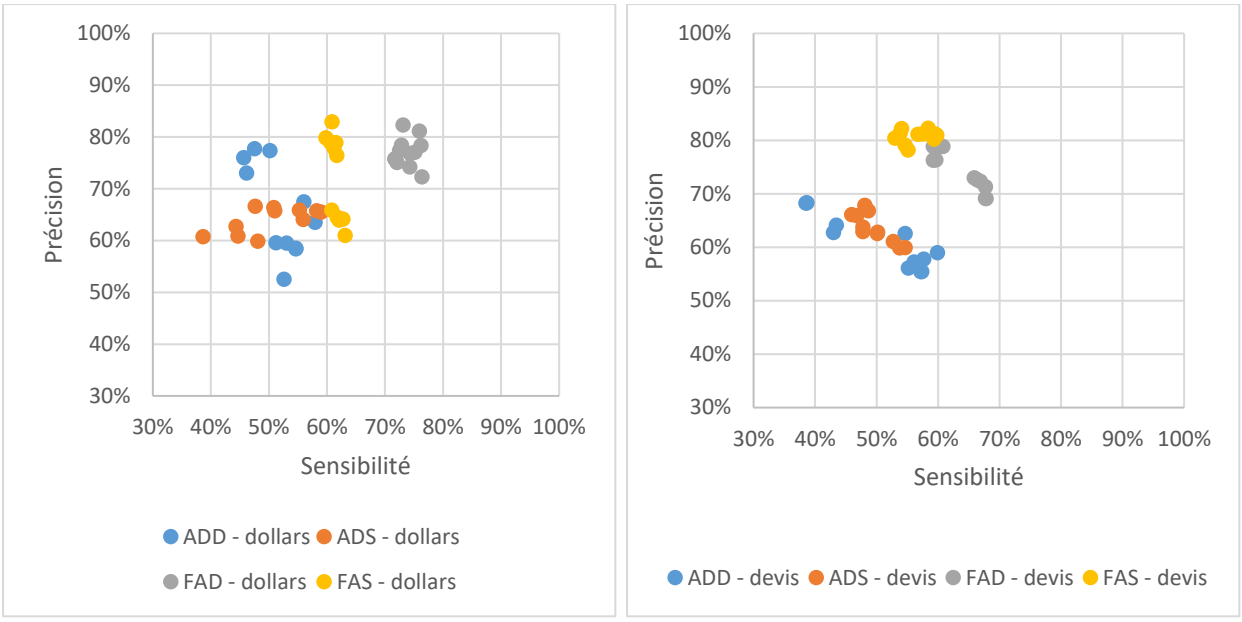

Figure F.4 : Précision de la classe annulation en fonction de la sensibilité (en dollars à gauche et devis à droite) - Client 23

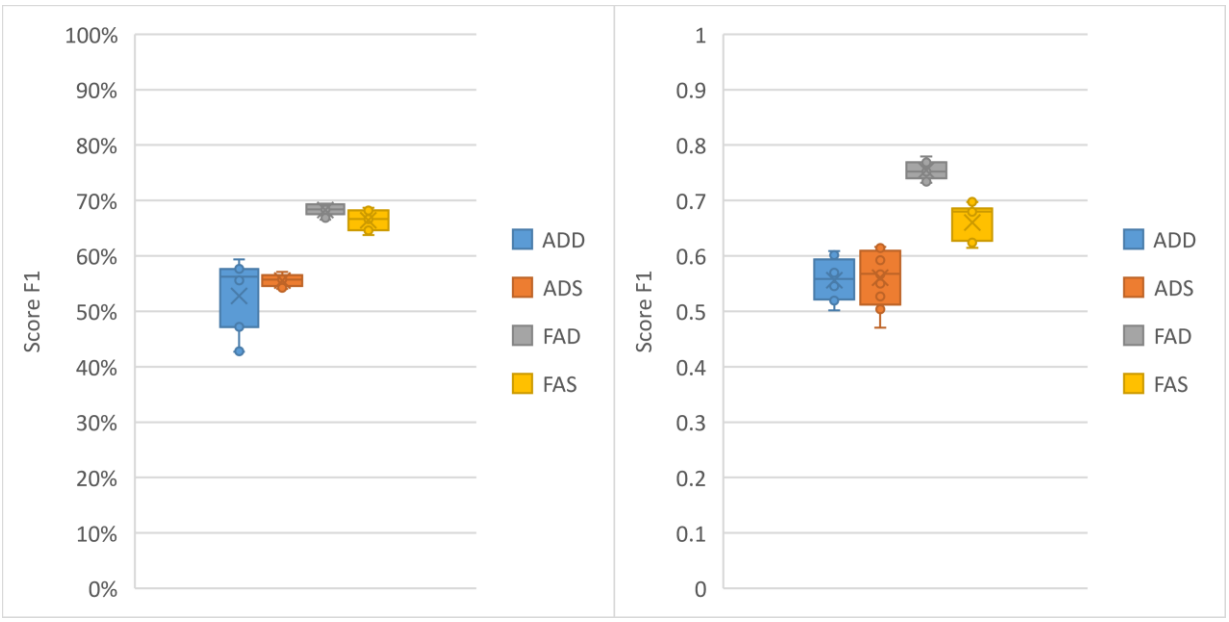

Figure F.5 : Score F1 en fonction du modèle et du regroupement pour la classe annulation (en dollars à gauche et devis à droite) – Client 23

Pour le client 23, les meilleurs modèles arrivent à récupérer 76 % des devis annulés avec 82 % de précision. (Figure F.4 à gauche). On retrouve bien dans le score F1 (Figure F.5) le bon score du FAD en devis.

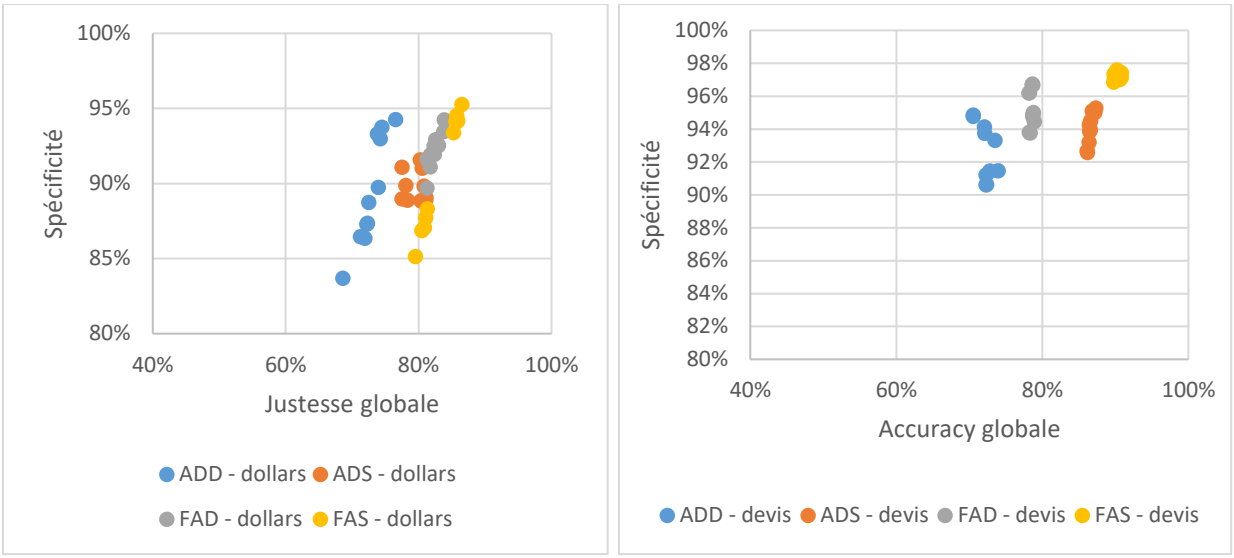

Figure F.6 : Spécificités de la classe annulation en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 23

La spécificité est très bonne malgré un nombre élevé de devis autres que Q0 pour les modèles FAD et FAS. (Figure F.6)

b. Classe vente et vente sur le bon quart

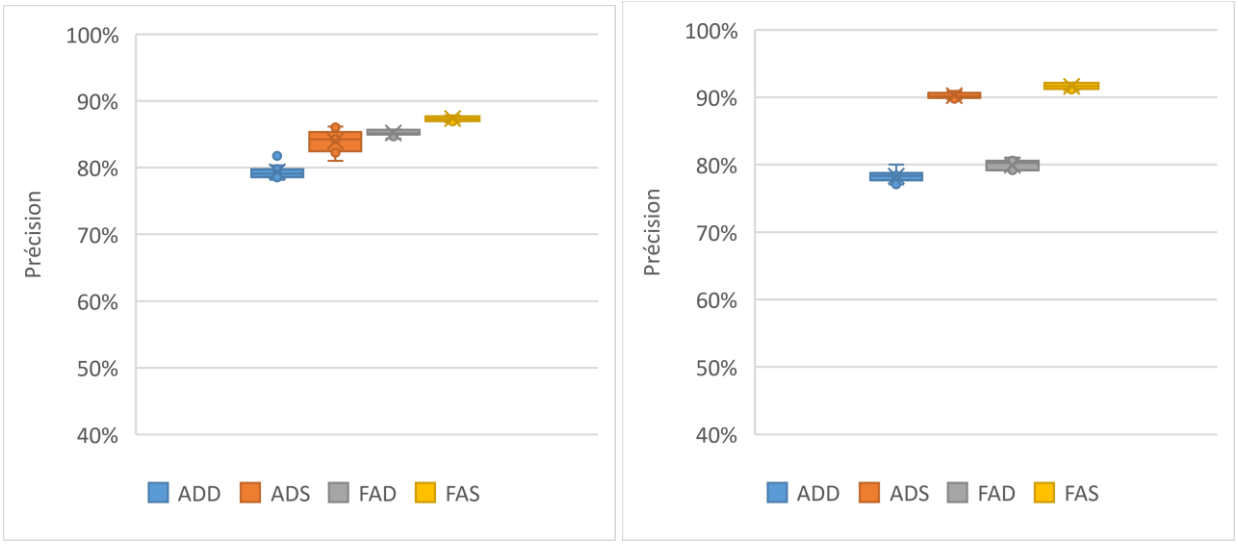

Figure F.7 : Précisions en fonction du modèle et du regroupement pour la classe vente ou Q0 (en dollars à gauche et devis à droite) - Client 23

La précision des modèles est correcte si on compare à la prévalence des devis « VENDU » ou Q0 avec un avantage pour les modèles simples si on ne prend pas en compte le poids des devis avec un très faible impact des regroupements.

La justesse globale semble corrélée à la précision. L'augmentation de l'une améliore l'autre (Figure F.8). Les modèles simples sont meilleurs si on ne prend pas en compte le poids des devis.

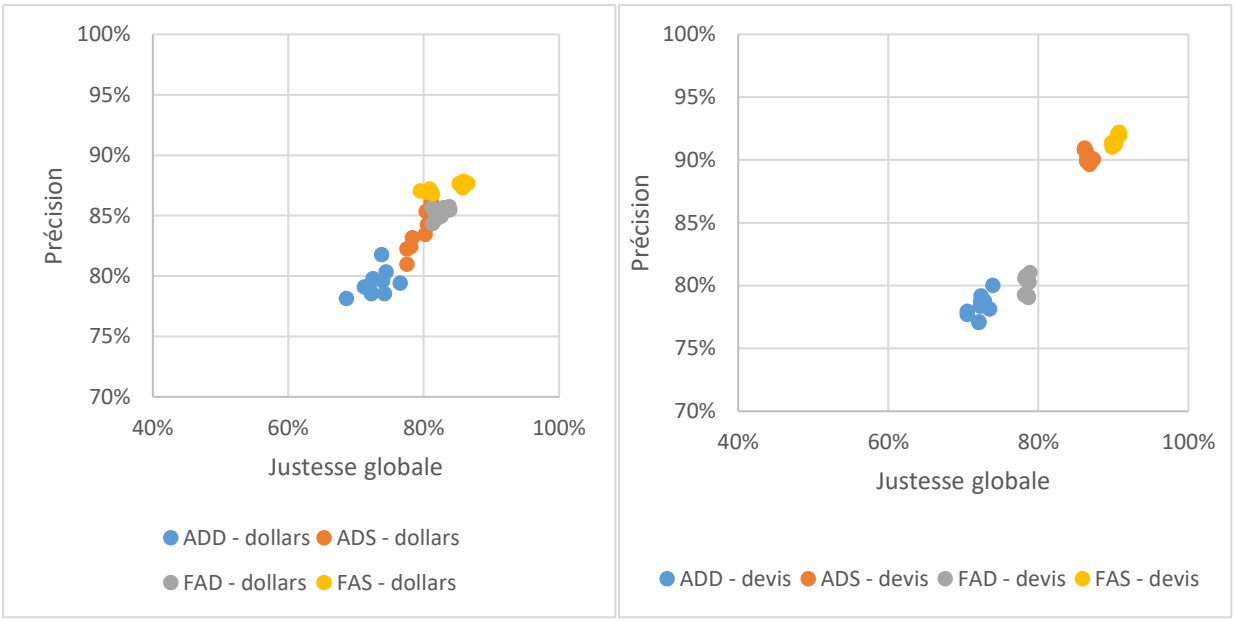

Figure F.8 : Précision de la classe « VENDU » ou Q0 en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 23

Pour le client 23, tous les modèles sont équivalents en termes de sensibilité et de précision pour la classe « VENDU » ou Q0 (Figure F.9). Les regroupements ont peu d'impact sur les résultats.

Les modèles simples sont meilleurs si on ne prend pas en compte le poids des devis aussi pour la précision/sensibilité. (Figure F.9)

Figure F.9 : Précision de la classe « VENDU » ou Q0 en fonction de la sensibilité (en dollars à gauche et devis à droite) - Client 23

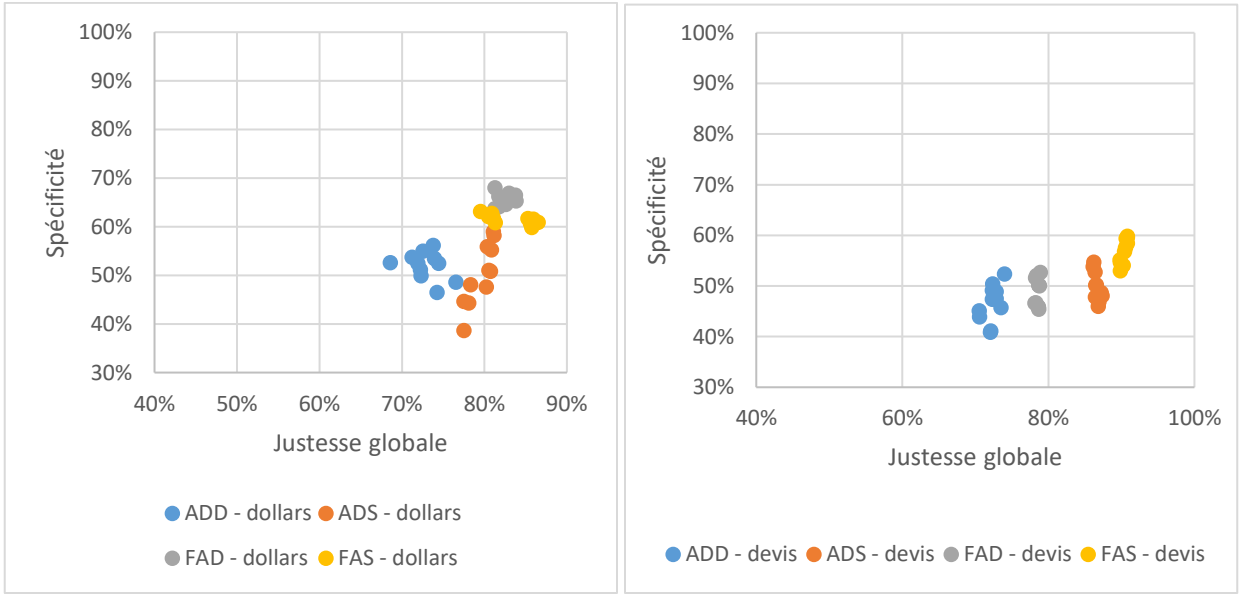

Figure F.10 : Spécificité de la classe « VENDU » ou « Q0 » en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 23

La spécificité de la classe Q0 ou « VENDU » est assez faible. Beaucoup de devis sont prédits en Q0 alors qu'ils ne le sont pas. Les modèles rencontrent des difficultés à bien classer les devis.

ANNEXE G RESULTATS CLIENT 93

1. Justesse globale

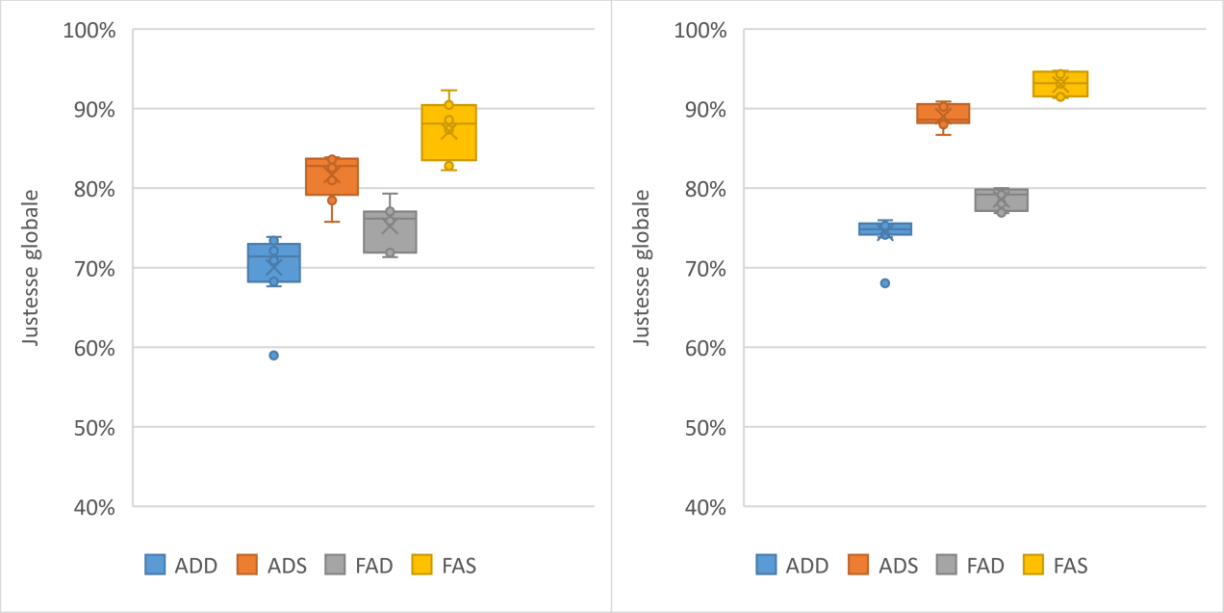

Figure G.1 : Justesse globale (en dollars à gauche et devis à droite) – Client 93

La justesse globale est peu impactée par les regroupements. Les résultats sont très similaires dans leurs types (simple ou détaillé) (Figure G.1).

8. Précision, sensibilité, spécificité, F1-score et justesse globale

a. Classe annulation

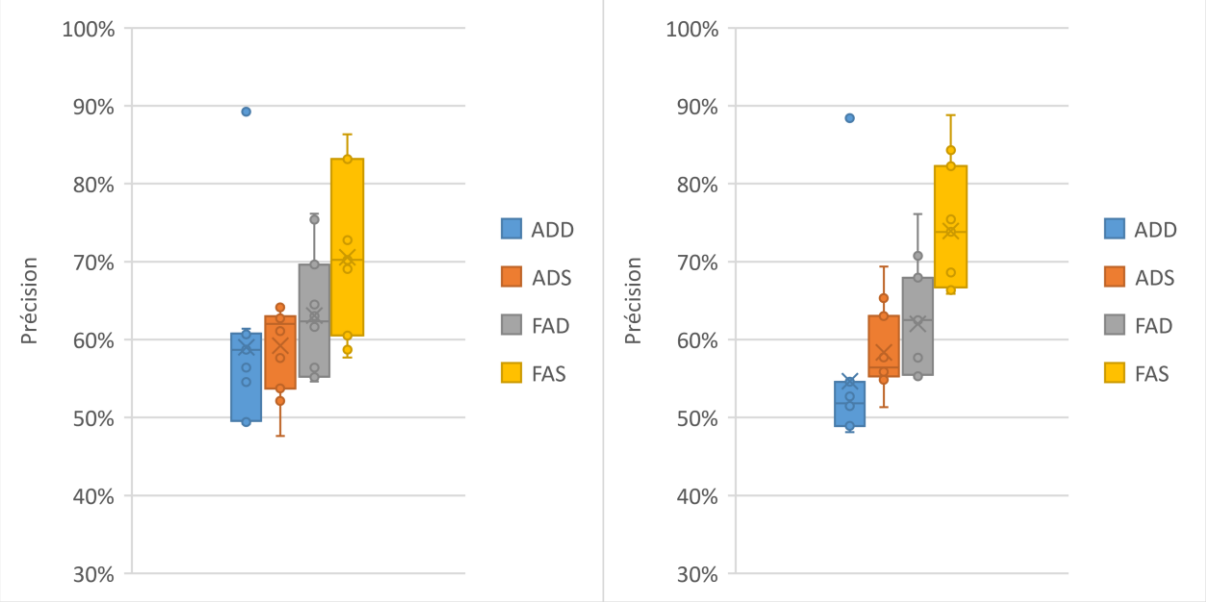

Figure G.2 : Précision en fonction du modèle et du regroupement pour la classe annulation (en dollars à gauche et devis à droite) – Client 93

La précision est très impactée par les regroupements (Figure G.2). Les résultats ne sont pas toujours très bons pour ce client avec comme meilleure précision un score de 86% en dollars par le modèle FAS.

Figure G.3: Précision de la classe annulation en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 93

Figure G.4 : Précision de la classe annulation en fonction de la sensibilité (en dollars à gauche et devis à droite) - Client 93

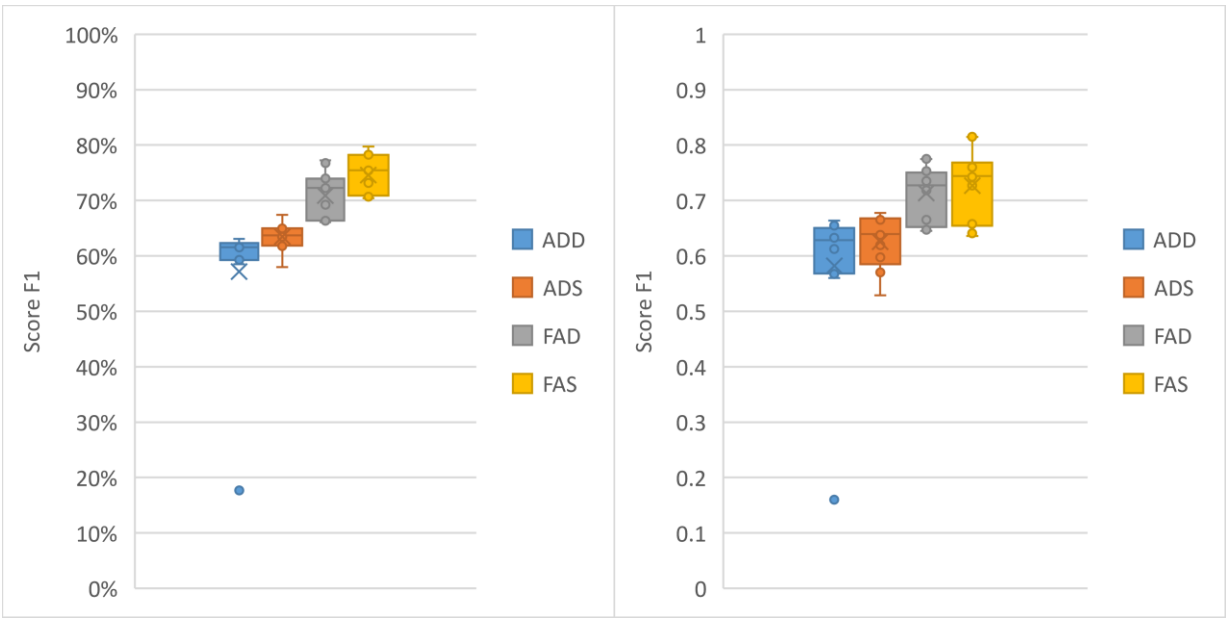

Figure G.5 : Score F1 en fonction du modèle et du regroupement pour la classe annulation (en dollars à gauche et devis à droite) – Client 93

Pour le client 93, les meilleurs modèles arrivent à récupérer 76 % des devis annulés avec 84 % de précision. (Figure G.4 à droite). On retrouve bien dans le score F1 (Figure G.5) de grande différence entre les regroupements.

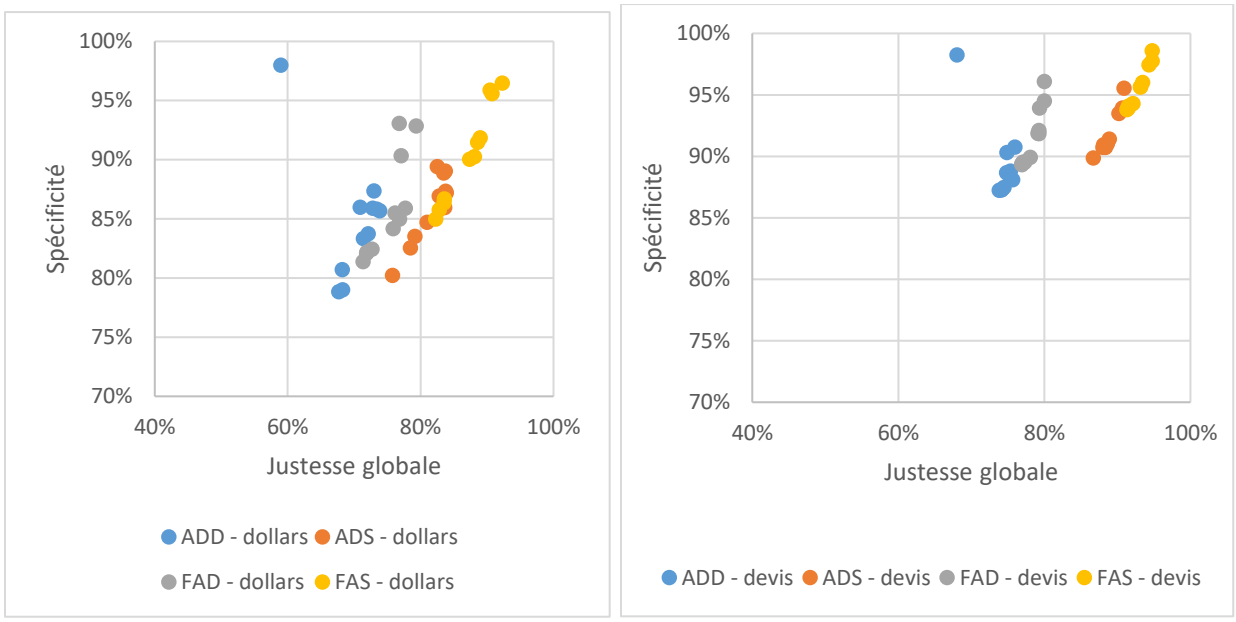

Figure G.6 : Spécificités de la classe annulation en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 93

La spécificité est très bonne, car la majorité des devis sont vendus et bien classés en vendu. (Figure G.6)

b. Classe vente et vente sur le bon quart

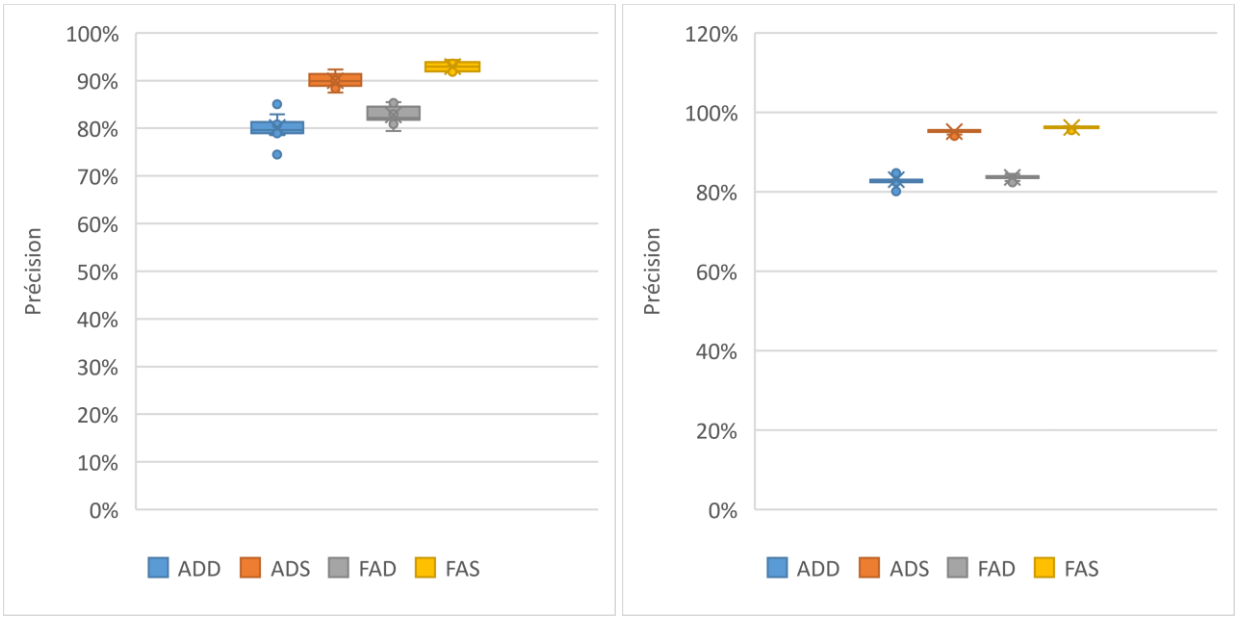

Figure G.7 : Précisions en fonction du modèle et du regroupement pour la classe vente ou Q0 (en dollars à gauche et devis à droite) - Client 93

La précision des modèles est correcte si on compare à la prévalence des devis « VENDU » ou Q0 avec un avantage pour les modèles simples avec un très faible impact des regroupements.

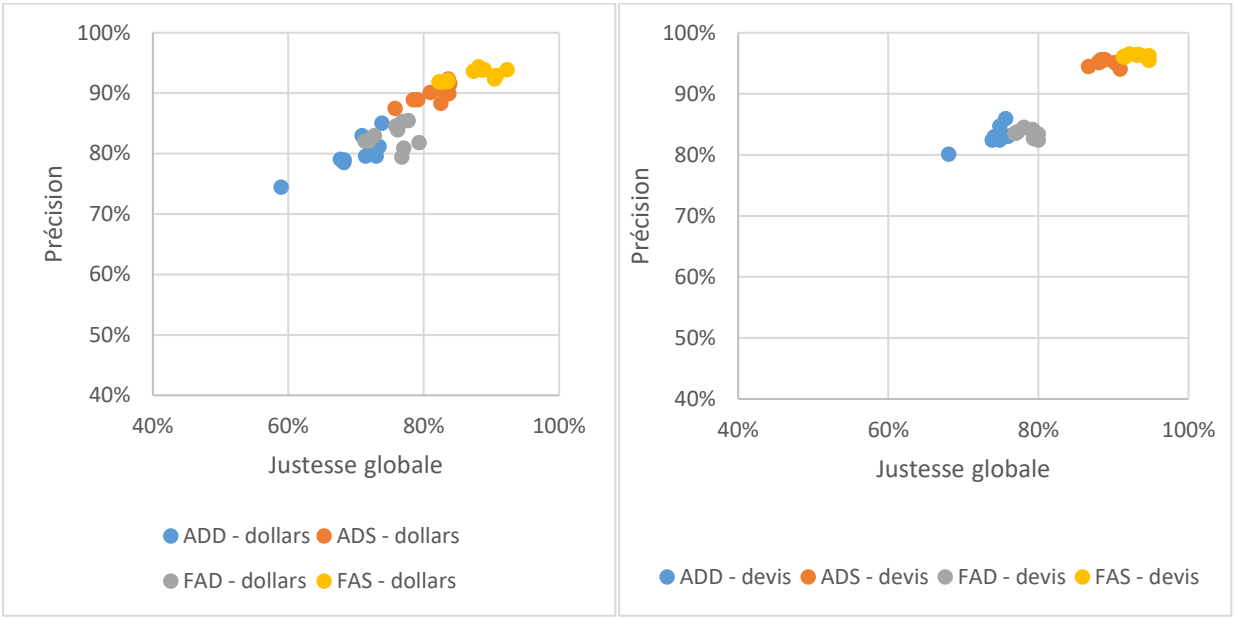

Figure G.8 : Précision de la classe « VENDU » ou Q0 en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 93

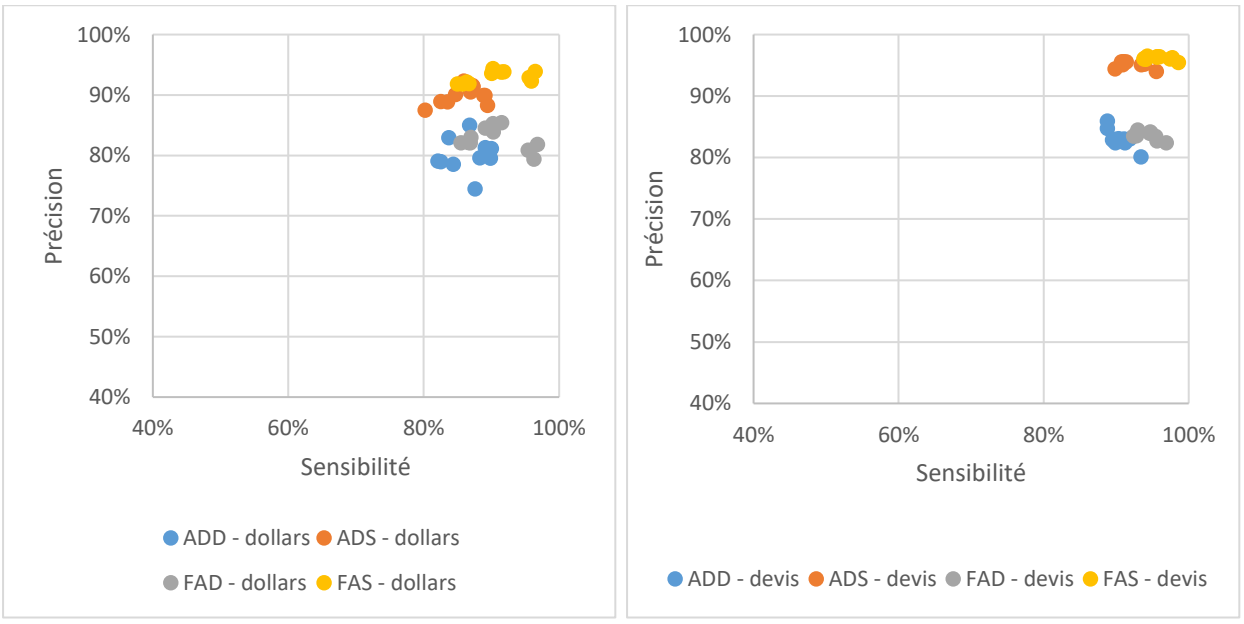

Figure G.9 : Précision de la classe « VENDU » ou Q0 en fonction de la sensibilité (en dollars à gauche et devis à droite) - Client 93

Pour le client 93, tous les modèles sont équivalents en termes de sensibilité si on prend les catégories simple ou détaillée (Figure G.9). Les modèles FAS sont cependant supérieurs.

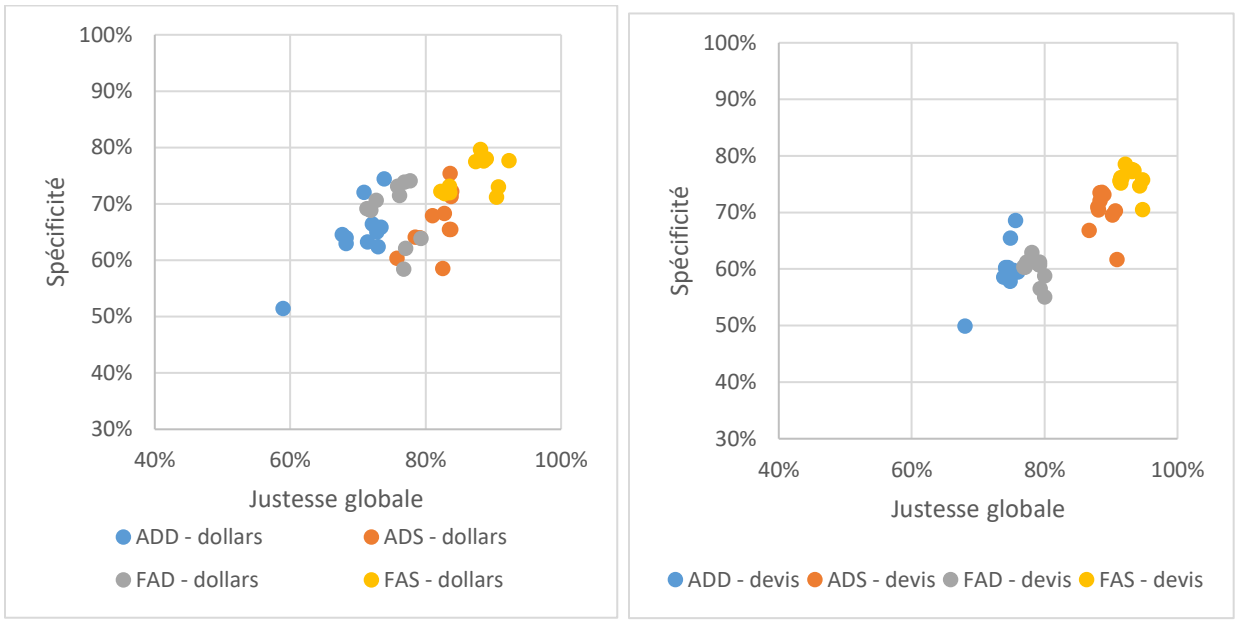

Figure G.10 : Spécificité de la classe « VENDU » ou « Q0 » en fonction de la justesse globale (en dollars à gauche et devis à droite) - Client 93

La spécificité de la classe Q0 ou « VENDU » est moyenne, car beaucoup de devis sont encore prédits en Q0 alors qu'ils n'appartiennent pas à cette classe. On arrive cependant à de très bons scores pour le modèle FAS en termes de dollars (Figure G.10).

ANNEXE H IMPORTANCE DES VARIABLES

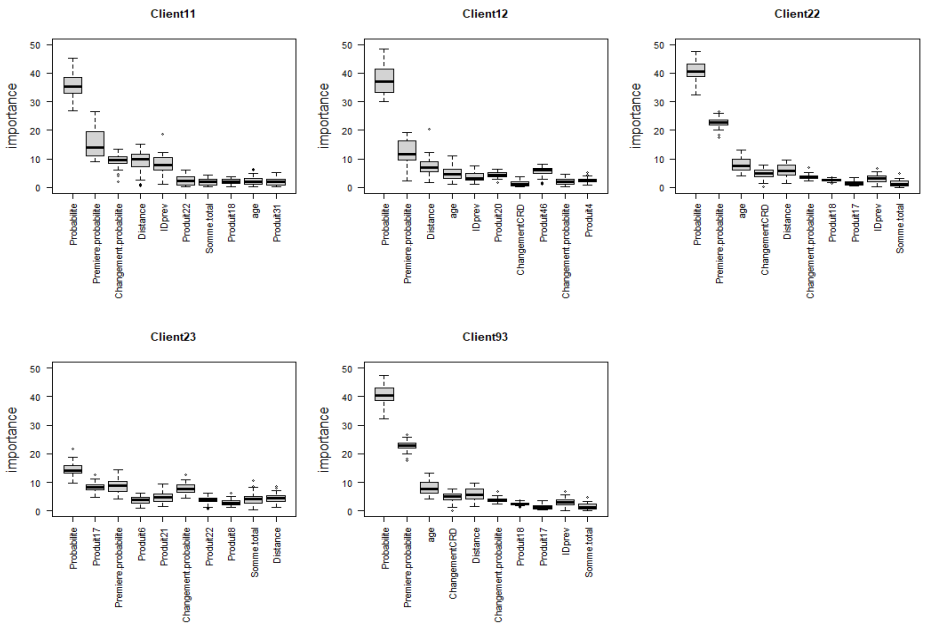

Figure H.1 : Importance des variables pour les ADD pour les 5 clients

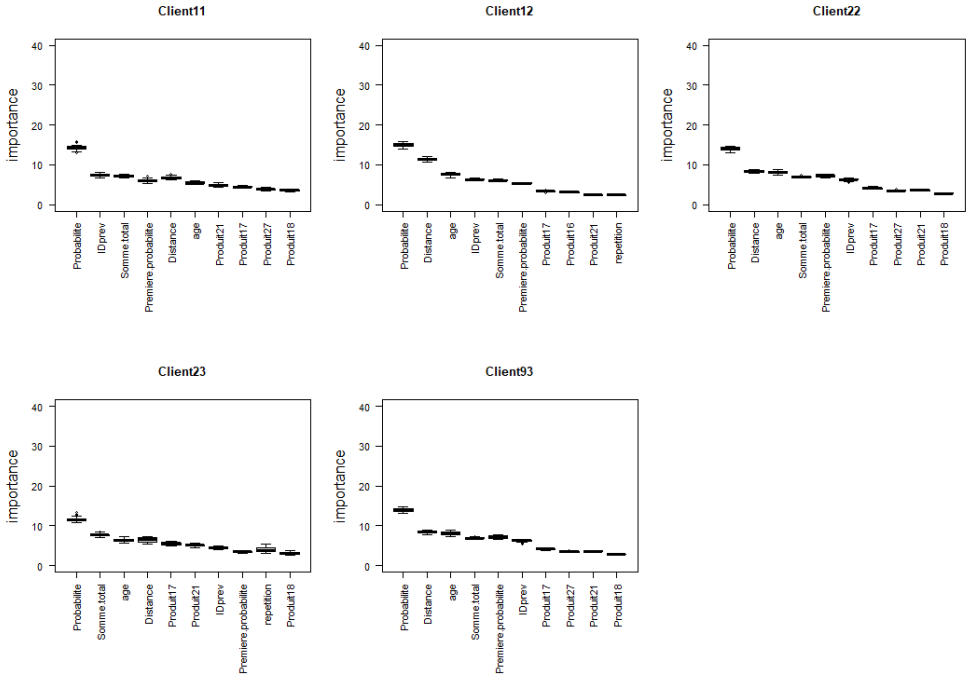

Figure H.2 : Importance des variables pour les FAD pour les 5 clients

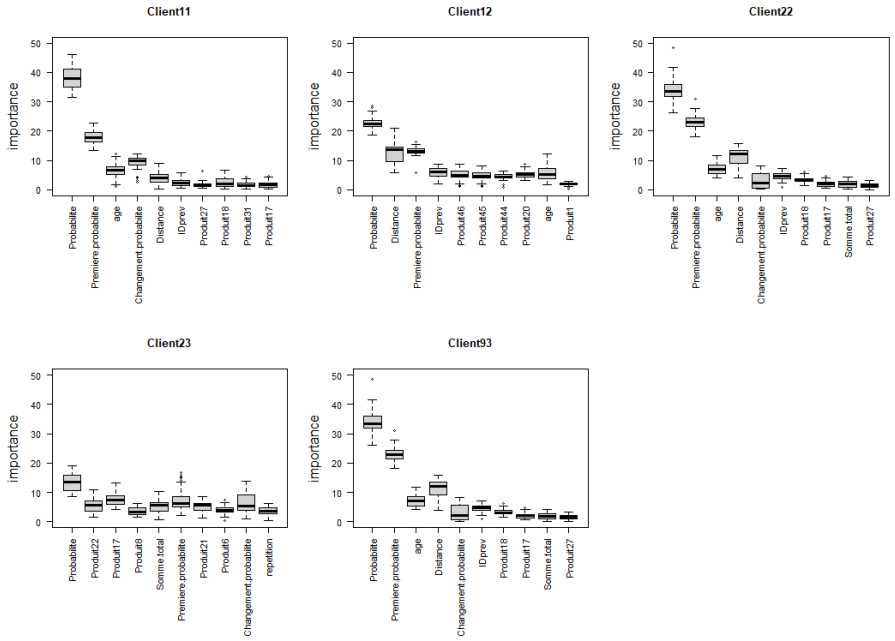

Figure H.3 : Importance des variables pour les ADS pour les 5 clients

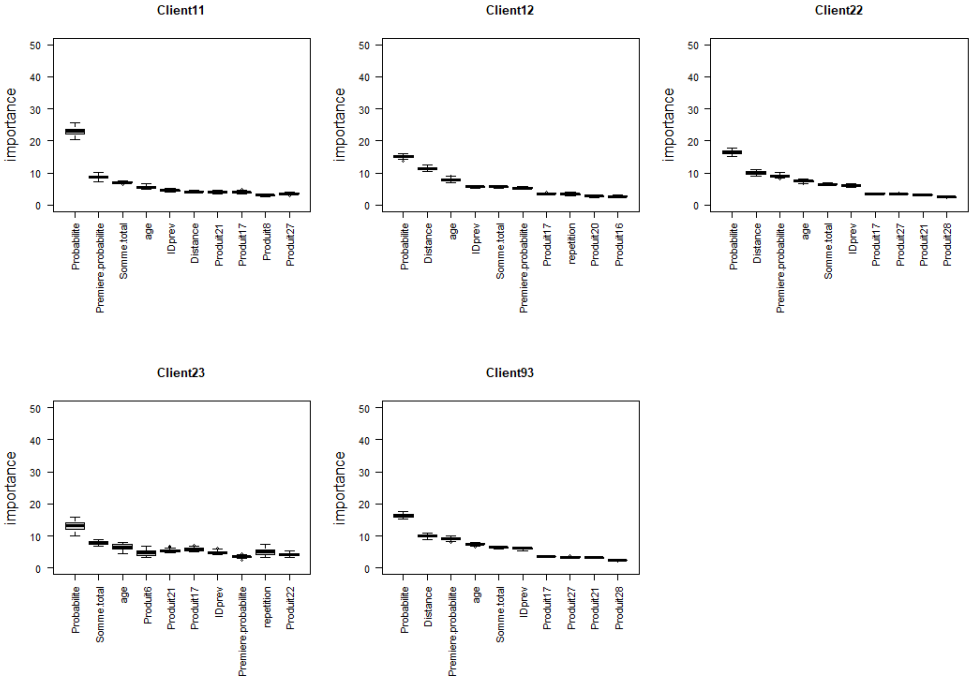

Figure H.4 : Importance des variables pour les FAS pour les 5 clients