UCLouvain
FUCaM MONS

LOUVAIN
School of Management

Louvain School of Management

Prédiction du churn client par extraction sémantique structurée : une comparaison empirique des embeddings contextuels et du pipeline RAG

Auteur : Nanat Inas
Promoteur(s) : Corentin Vande Kerckhove
Année académique : 2025-2026
Master [120] Ingénieur de gestion, à finalité spécialisée

Déclaration relative à l'utilisation d'outils d'intelligence artificielle

Nous reconnaissons que les outils d'IA peuvent constituer une aide précieuse dans le cadre de la rédaction d'un mémoire, mais ils ne sont pas infaillibles. La transparence favorise la confiance, et la reconnaissance du rôle de l'IA renforce la crédibilité du travail. En conséquence, l'utilisation de tout outil d'IA dans ce mémoire a été guidée par les principes suivants : (1) **évaluation critique** des productions générées, toute correction étant effectuée sur la base de l'expertise de l'auteur ; (2) **transparence** quant à l'utilisation de Claude (Anthropic), ChatGPT (OpenAI) et NotebookLM (Google), qui ont contribué au travail sans se substituer au jugement humain ; (3) **responsabilité éthique** dans l'identification des biais potentiels introduits par ces outils.

Dans le cadre de la préparation de ce mémoire, l'auteur a mobilisé ces outils pour les usages suivants :

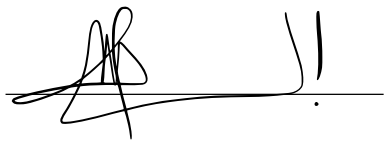
— **Aide à la rédaction :** ChatGPT (OpenAI) a été utilisé pour reformuler et paraphraser des passages écrits afin d'améliorer la clarté et le registre académique.

— **Support à la revue de littérature :** NotebookLM (Google) a été utilisé pour soutenir la consultation et la synthèse de sources académiques durant la phase de revue de littérature.

— **Aide au code et débogage :** Claude (Anthropic) a été utilisé pour assister la structuration, l'optimisation et le débugage du code Python développé pour le pipeline de données et les analyses empiriques.

— **Aide à la mise en page LaTeX :** Claude (Anthropic) a été utilisé pour assister la mise en page et la mise en forme du document de mémoire en LaTeX, notamment pour la syntaxe des expressions mathématiques, des tableaux, des figures et l'intégration bibliographique.

Après chaque utilisation, l'auteur a relu, vérifié et édité le contenu produit. L'auteur assume l'entière responsabilité du résultat final, affirmant que celui-ci reflète un travail intellectuel original augmenté par un usage ciblé et responsable de l'IA.

Date : 27/05/2026 Signature : 

Résumé

Dans un contexte de marché saturé, la prédiction du churn est devenue un enjeu stratégique majeur pour les entreprises. Si les modèles prédictifs reposent traditionnellement sur des données transactionnelles structurées, les avis textuels laissés par les clients après leur expérience d'achat constituent une source d'information complémentaire encore largement sous-exploitée. Ce mémoire s'interroge sur la forme optimale de représentation de ces données textuelles dans un cadre prédictif, en posant la question suivante : dans quelle mesure une extraction sémantique structurée via un pipeline RAG améliore-t-elle la prédiction du churn client par rapport aux données structurées seules ou aux embeddings contextuels ?

Pour y répondre, quatre conditions expérimentales sont comparées sur le Brazilian E-Commerce Public Dataset d'Olist, qui combine à l'échelle transactionnelle des données structurées et des avis textuels en portugais. La première condition repose exclusivement sur des variables transactionnelles. La deuxième intègre des embeddings contextuels produits par un modèle SentenceTransformer multilingue. La troisième mobilise un pipeline de type Retrieval-Augmented Generation (RAG), qui structure l'information textuelle autour de quatre dimensions théoriques du churn identifiées dans la littérature, à savoir la livraison, la qualité produit, le service client et le prix, afin de produire des variables sémantiques interprétables. Enfin, la dernière condition constitue une référence non supervisée fondée sur une approche zero-shot avec GPT-4o mini. Le dispositif repose sur un split temporel, des tests de permutation et une analyse bootstrap garantissant la robustesse des comparaisons.

Les résultats montrent un apport massif et statistiquement significatif des données textuelles par rapport aux données structurées seules. À pleine capacité d'entraînement, les embeddings contextuels et le pipeline RAG atteignent des performances équivalentes. Toutefois, à volume d'entraînement réduit, le RAG surpasse significativement les embeddings, indiquant une meilleure efficience par observation. Au-delà de la performance, le RAG produit des variables directement interprétables et actionnables, là où les embeddings fonctionnent comme une boîte noire. Quelle que soit la méthode mobilisée, la livraison émerge comme le déterminant principal du churn dans Olist, résultat cohérent avec la littérature sur le e-commerce.

Ce mémoire explore ainsi un usage du RAG distinct de son application classique en questions-réponses, en l'utilisant comme générateur de variables structurées destinées à alimenter un modèle prédictif supervisé.

Remerciements

Ce travail touchant à sa fin, il est temps d'exprimer ma gratitude à toutes les personnes qui ont contribué à sa réalisation.

Je souhaite tout d'abord remercier chaleureusement mon promoteur, Monsieur Corentin Vande Kerckhove, pour son accompagnement tout au long de ce mémoire. Il m'a guidée dès les premières étapes, de la définition de la question de recherche jusqu'à la finalisation de ce travail, et a toujours su m'orienter avec pertinence lorsque cela était nécessaire. Sa disponibilité constante, malgré mes nombreuses sollicitations, a été précieuse et ses conseils ont grandement contribué à l'aboutissement de ce mémoire.

Une pensée particulière va ensuite à mes parents, qui m'ont soutenue et encouragée sans relâche depuis le début de mes études universitaires, et même bien avant encore. Leur confiance en moi, leur présence constante ainsi que toutes les formes d'aide qu'ils m'ont apportées au fil des années ont joué un rôle essentiel dans mon parcours.

Je voudrais aussi remercier mon copain, qui a toujours été présent pour moi, et plus particulièrement durant cette période intense du mémoire. Il m'a soutenue dans les moments de doute, durant les longues périodes de travail , ainsi que lors des nombreuses remises en question qui ont accompagné sa réalisation. Sa présence, son écoute et sa capacité à m'aider à prendre du recul lorsque j'en avais besoin ont représenté un véritable soutien.

Mes remerciements vont également à mes amis, avec qui j'ai partagé un quadrimestre particulièrement intense. Entre le stress du mémoire, celui du stage, les sessions de travail et les moments de découragement, nous avons tous vécu cette période intense ensemble. Pouvoir partager ces difficultés avec des personnes traversant les mêmes étapes a rendu ce quadrimestre plus supportable. Au-delà de ce mémoire, ces cinq années passées ensemble ont créé des liens forts, et il est émouvant de voir cette étape de notre vie arriver à son terme après un parcours vécu côte à côte.

Enfin, je souhaite remercier toutes les personnes qui, de près ou de loin, ont contribué à l'élaboration de ce mémoire. Même si elles ne peuvent être citées individuellement, leur soutien n'en reste pas moins précieux et sincèrement apprécié.

Table des matières

Déclaration relative à l'utilisation d'outils d'IA i

Résumé ii

Remerciements iii

Liste des figures vii

Liste des tableaux viii

Liste des annexes ix

Liste des abréviations x

1 Introduction 1

2 Revue de littérature 3

2.1 Le churn en e-commerce : enjeux et déterminants . . . 3

2.1.1 Enjeux business de la prédiction client . . . 3

2.1.2 Définition et typologies du churn . . . 3

2.1.3 Déterminants du churn en e-commerce . . . 4

2.2 Les données pour la prédiction du churn . . . 5

2.2.1 Données structurées CRM et leurs limites . . . 5

2.2.2 Données textuelles et voix du client . . . 6

2.3 Modèles prédictifs pour la classification du churn . . . 7

2.3.1 Cadre général du machine learning supervisé . . . 7

2.3.2 Modèles pour des données tabulaires déséquilibrées . . . 8

2.4 Représentation du texte pour l'analyse prédictive . . . 9

2.4.1 Approches creuses et les embeddings statiques . . . 9

2.4.2 Embeddings contextuels : Transformers, BERT et SentenceTransformer . . . 9

2.4.3 LLM génératifs et architecture RAG . . . 10

2.4.4 Positionnement du mémoire et contribution scientifique . . . 11

3 Méthodologie 12

3.1 Positionnement épistémologique et design de recherche . . . 12

3.2 Données . . . 13

3.2.1 Présentation du *dataset* . . . 13

3.2.2 Construction de la variable cible . . . 14

3.2.3 Construction du dataset commun et split temporel . . . 15

3.3 Conditions expérimentales . . . 16
3.3.1 Condition M1 : Données structurées seules . . . 16
3.3.2 Condition M2 : Données structurées + *embeddings* contextuels . . . 16
3.3.3 Condition M3 : Données structurées + pipeline *RAG* . . . 17
3.3.4 Condition M4 : LLM zero-shot . . . 20
3.4 Modèles et optimisation des hyperparamètres . . . 20
3.5 Évaluation . . . 21
3.6 Conclusion du chapitre . . . 22

4 Résultats et Analyses **23**
4.1 Paramètres empiriques du pipeline RAG . . . 23
4.2 Performance globale des conditions expérimentales . . . 23
4.3 Comparaison à taille d'entraînement fixe entre embeddings et RAG . . . 25
4.4 Diagnostic d'overfitting . . . 26
4.5 Feature importance . . . 26
4.6 Analyses complémentaires . . . 28
4.6.1 Calibration . . . 28
4.6.2 Courbe de gain cumulatif . . . 29
4.7 Analyse de robustesse . . . 30
4.8 Conclusion du chapitre . . . 31

5 Discussion **32**
5.1 Interprétation des résultats au regard de la question de recherche . . . 32
5.2 Le signal de churn dans Olist . . . 33
5.3 Le zero-shot comme référence non supervisée . . . 34
5.4 Implications managériales . . . 34
5.5 Considérations éthiques . . . 36
5.6 Limites . . . 37
5.7 Pistes de recherches futures . . . 38
5.8 Conclusion du chapitre . . . 39

6 Conclusion Générale **40**

Bibliographie **41**

Annexes **44**
1 Distribution des notes de review sur le jeu d'entraînement . . . 44
2 Liste complète des features M1 . . . 44
3 Base de connaissances du pipeline RAG . . . 45
4 Learning curve M2 . . . 51
5 Analyse de sensibilité k (FAISS) . . . 52

6 Learning curve M3 52
7 Corrélations scores RAG vs note de satisfaction 53
8 Grilles d'hyperparamètres 53
9 Code source 55
10 Résultats finaux complets sur le jeu de test 55
11 Feature importance complète M1 XGB et M3 XGB 56
12 Analyse de robustesse (Régression) 58

Liste des figures

1 Architecture du pipeline RAG 18

2 Feature importance – M1 XGB 27
3 Feature importance – M3 XGB 27
4 Courbes de calibration des quatre conditions expérimentales 28
5 Courbes de gain cumulatif 29

6 Distribution des notes de review 44
7 Learning curve – Condition M2 51
8 Learning curve – Condition M3 52

Liste des tableaux

1 Synthèse des quatre conditions expérimentales 13

2 Meilleurs modèles par condition : performances sur le jeu de test 24
3 Tests de permutation 24
4 Comparaison de M2 et M3 à taille d'entraînement fixe et jeu de test complet 25
5 Proportion de churners capturés selon le seuil de ciblage 30
6 Analyse de robustesse 30

7 Liste complète des features M1 44
12 Analyse de sensibilité k (FAISS) 52
13 Corrélations scores RAG vs note de satisfaction 53
14 Moyenne des scores RAG par note de satisfaction 53
15 Grille d'hyperparamètres – Régression Logistique 53
16 Grille d'hyperparamètres – XGBoost M1 et M2 54
17 Grille d'hyperparamètres – XGBoost M3 54
18 Résultats finaux complets sur le jeu de test 55
19 Feature importance complète – M1 XGB 56
20 Feature importance complète – M3 XGB 57
21 Métriques par condition – Analyse de robustesse 58
22 Tests de Wilcoxon – Significativité des différences 58

Liste des annexes

— Annexe 1 : Distribution des notes de review sur le jeu d'entraînement

— Annexe 2 : Liste complète des features M1

— Annexe 3 : Base de connaissances du pipeline RAG

— Annexe 4 : Learning curve M2

— Annexe 5 : Analyse de sensibilité k (FAISS)

— Annexe 6 : Learning curve M3

— Annexe 7 : Corrélations scores RAG vs note de satisfaction

— Annexe 8 : Grilles d'hyperparamètres

— Annexe 9 : Code source

— Annexe 10 : Résultats finaux complets sur le jeu de test

— Annexe 11 : Feature importance complète M1 XGB et M3 XGB

— Annexe 12 : Analyse de robustesse (Régression)

Liste des abréviations

Abréviation	Définition
LR	Régression Logistique (Logistic Regression)
XGB	Extreme Gradient Boosting (XGBoost)
RAG	Génération Augmentée par Récupération (Retrieval-Augmented Generation)
NLP	Traitement Automatique du Langage Naturel (Natural Language Processing)
LLM	Grand Modèle de Langage (Large Language Model)
PR-AUC	Aire sous la courbe Précision-Rappel (Precision-Recall Area Under the Curve)
ROC-AUC	Aire sous la courbe ROC (Receiver Operating Characteristic Area Under the Curve)
RFM	Récence, Fréquence, Montant
CRM	Gestion de la Relation Client (Customer Relationship Management)
CLV	Valeur Vie Client (Customer Lifetime Value)
FAISS	Facebook AI Similarity Search
IC	Intervalle de Confiance
MAE	Erreur Absolue Moyenne (Mean Absolute Error)
RMSE	Racine de l'Erreur Quadratique Moyenne (Root Mean Squared Error)
VoC	Voix du Client

Introduction

Dans un marché saturé, conserver un client existant coûte entre cinq et vingt-cinq fois moins cher que d'en acquérir un nouveau (Gupta & Zeithaml, 2006). Ce seul chiffre suffit à expliquer pourquoi la prédiction du churn est devenue un enjeu stratégique central pour les entreprises. Le churn, défini comme la tendance des clients à cesser leurs relations commerciales avec une entreprise (Figalist et al., 2019), réduit directement la valeur vie client et, par extension, la rentabilité à long terme. Dans des contextes non contractuels comme le e-commerce, où le départ d'un client n'est jamais formellement observable, identifier les signaux précurseurs de désengagement avant qu'ils ne se traduisent dans les comportements d'achat constitue un avantage compétitif décisif.

Chaque jour, des millions de consommateurs laissent des traces textuelles de leur expérience d'achat en ligne, que ce soit une note, un commentaire, parfois quelques mots griffonnés après une livraison ratée ou un produit décevant. Dans une récente méta-analyse portant sur 60 études et plus de 29 000 observations, Ghosh (2024) démontre que le bouche-à-oreille électronique (e-WOM), qui inclut les avis en ligne, exerce un impact significatif sur les attitudes et les intentions d'achat des consommateurs. Pourtant, cette même donnée abondante et chargée d'informations comportementales reste largement sous-exploitée dans les modèles prédictifs de rétention client (Balducci & Marinova, 2018 ; De Caigny et al., 2020).

Ce paradoxe est d'autant plus frappant dans le contexte du e-commerce brésilien, marché ayant atteint 235 milliards de reais en 2025 (Ma, 2026), où le volume de feedback client généré dépasse largement toute capacité d'analyse manuelle. Pourtant, les modèles de prédiction du churn reposent encore majoritairement sur des données transactionnelles structurées, laissant de côté l'information contenue dans les avis textuels (De Caigny et al., 2020).

Si les données textuelles apparaissent comme une source d'information prometteuse pour mieux comprendre le comportement client, la question de leur représentation optimale dans un cadre prédictif reste ouverte. Les embeddings contextuels, les pipelines de type Retrieval-Augmented Generation (RAG) et les modèles de langage zero-shot reposent sur des logiques d'extraction sémantique différentes, dont l'impact comparatif sur la prédiction du churn n'a pas encore fait l'objet d'une évaluation empirique rigoureuse. Ce mémoire se propose de répondre à cette question à travers l'interrogation suivante :

> ***Dans quelle mesure une extraction sémantique structurée via un pipeline RAG améliore-t-elle la prédiction du churn client par rapport aux données structurées seules ou aux embeddings contextuels ?***

Cette question implique non seulement l'évaluation de la performance prédictive de chaque approche mais également leur capacité à produire des représentations exploitables dans une logique analytique et métier.

Pour y répondre, quatre conditions expérimentales sont comparées sur le Brazilian E-Commerce Public Dataset d'Olist : données structurées seules, embeddings contextuels, pipeline RAG et LLM zero-shot. Ce design permet d'isoler l'effet propre de chaque forme de représentation textuelle sur la performance prédictive.

Dans ce mémoire, nous poserons d'abord le cadre théorique à travers une revue de littérature, avant de détailler la démarche empirique utilisée. Les deux dernières parties présentent les résultats et les discutent sous différents angles, à savoir la performance, l'interprétabilité ainsi que les implications managériales.

Revue de littérature

2.1 Le churn en e-commerce : enjeux et déterminants

2.1.1 Enjeux business de la prédiction client

La satisfaction client correspond au jugement du consommateur selon lequel un produit ou un service répond, ou non, à ses attentes (Gupta & Zeithaml, 2006). Elle se traduit directement en valeur financière pour l'entreprise, mesurée à travers la Valeur Vie Client (Customer Lifetime Value, CLV), définie comme la valeur actuelle des futurs profits générés par un client sur l'ensemble de sa relation avec l'entreprise. Plusieurs études montrent qu'une amélioration même marginale des taux de rétention peut avoir un impact significatif sur cette valeur. Une augmentation de seulement 5 % de la rétention client peut ainsi entraîner une hausse des profits, comprise entre 25 % et 85 % (Gupta & Zeithaml, 2006).

Dans ce contexte, adopter une stratégie non ciblée ou réactive, tel qu'offrir des avantages ou des réductions à l'ensemble de la base client, s'avère être inefficace et coûteux (Jahromi et al., 2014). Identifier les clients à risque de manière proactive devient alors un enjeu central. Les modèles de prédiction du churn répondent précisément à cet objectif en permettant d'anticiper le départ d'un client avant qu'il ne se concrétise, et ainsi d'orienter les actions de rétention vers ceux pour qui elles ont le plus de sens (Manzoor et al., 2024).

2.1.2 Définition et typologies du churn

Le churn, ou attrition client, désigne la tendance des clients à cesser leurs relations commerciales avec une entreprise pour se tourner vers d'autres fournisseurs (Figalist et al., 2019). La manière dont ce phénomène est identifié dépend directement du mode de fonctionnement de l'entreprise.

Dans les environnements contractuels, le départ du client est clairement observable puisque le client rompt son contrat ou ne le renouvelle pas à son terme (Fader & Hardie, 2009). L'attrition étant directement mesurable, l'enjeu prédictif porte plutôt sur la durée de vie résiduelle du client que sur la détection de son départ (Fader & Hardie, 2009).

Dans les environnements non-contractuels, son identification devient plus complexe. Les clients quittent l'entreprise de manière silencieuse, sans signal explicite (Fader & Hardie, 2009 ; Jahromi et al., 2014). C'est notamment le cas dans le commerce de détail, le e-commerce, ou encore la vente par correspondance (Fader & Hardie, 2009). Le départ n'étant pas directement observable, sa définition devient alors arbitraire et dépend du contexte sectoriel. Il est généralement défini à partir d'une période d'inactivité, en observant l'absence d'achat sur une fenêtre temporelle donnée

(Jahromi et al., 2014 ; Manzoor et al., 2024).

Dès lors, l'anticipation du churn repose nécessairement sur des proxies, puisque le départ n'est jamais directement observable (Fader & Hardie, 2009). Les proxies comportementaux constituent le principal levier de prédiction, en s'appuyant sur les données transactionnelles et les historiques d'usage afin de capturer les comportements réels des clients (Buckinx & Van den Poel, 2005 ; Manzoor et al., 2024). À l'inverse, les proxies attitudinaux permettent d'aller au-delà du simple constat comportemental et de comprendre les causes du désengagement plutôt que de simplement l'observer (Manzoor et al., 2024). Ils sont particulièrement pertinents lorsque l'entreprise dispose de volumes importants de données textuelles non structurées, tels que des avis en ligne ou encore des réclamations post-achat, qui reflètent le ressenti du client au-delà des seules traces transactionnelles (Manzoor et al., 2024).

Ainsi, la compréhension du churn ne peut se limiter à sa définition opérationnelle, mais nécessite d'identifier les facteurs qui en expliquent l'apparition.

2.1.3 Déterminants du churn en e-commerce

Prédire le churn suppose d'abord de comprendre ce qui le provoque. Dans le contexte du e-commerce, plusieurs dimensions de l'expérience client ont été identifiées dans la littérature comme des déterminants majeurs du désengagement.

La livraison constitue la première de ces dimensions. Si la facilité d'utilisation des sites et la sécurité des paiements sont désormais largement considérées comme des acquis, les attentes des consommateurs se concentrent surtout sur la phase post-achat (Nguyen & Chanut, 2018). Il s'agit avant tout de livrer la commande correctement et dans les délais annoncés, toute défaillance à ce niveau ayant un impact direct sur la satisfaction globale et la relation client à long terme (Masyhuri, 2022). La traçabilité en temps réel s'est également imposée comme une norme, les clients attendant un suivi continu de leur commande. Un manque de visibilité génère ainsi de l'incertitude et dégrade la perception du service (Nguyen & Chanut, 2018). En cas de retards importants ou de réception d'un produit endommagé, la livraison devient un point de rupture susceptible d'affecter durablement la fidélité du client (Masyhuri, 2022). La logistique apparaît donc comme un levier central de rétention, dont les défaillances constituent un facteur direct de churn (Nguyen & Chanut, 2018 ; Archak et al., 2011).

La qualité du produit constitue un deuxième déterminant majeur du churn. Son impact s'explique par la théorie de la confirmation des attentes, selon laquelle le consommateur compare la performance réelle du produit à ses attentes initiales (Gupta & Zeithaml, 2006). Lorsque celle-ci est jugée insuffisante, l'intention de réachat diminue (Al-Ayed, 2022). Cet effet est d'autant plus marqué que la qualité recouvre plusieurs dimensions, notamment la fiabilité, la durabilité et la conformité à la description, sur lesquelles les clients fondent des attentes précises avant l'achat (Tirunillai & Tellis, 2014 ; De Caigny et al., 2020). L'écart entre la promesse et l'expérience vécue conditionne ainsi la probabilité de défection, d'autant que les clients insatisfaits ont tendance

à exprimer leur mécontentement publiquement, amplifiant l'impact pour l'entreprise (Masyhuri, 2022).

En l'absence d'interactions physiques, le service client constitue le principal point de contact humain de l'e-commerce, et ses défaillances figurent parmi les moteurs les plus directs du churn (Masyhuri, 2022). Les clients s'attendent à une prise en charge rapide, courtoise et compétente de leurs demandes. Dès lors, une incapacité à répondre dans des délais raisonnables ou à résoudre efficacement un problème génère une insatisfaction immédiate (Masyhuri, 2022 ; Nguyen & Chanut, 2018). Au-delà de la réactivité, la capacité de récupération après un incident, tel qu'un retard de livraison ou un défaut produit, constitue un facteur déterminant dans la décision de fidélisation ou de départ du client (Al-Ayed, 2022).

Le prix occupe une place paradoxale parmi les facteurs explicatifs du phénomène. Si plus de 60 % des clients citent la recherche du meilleur prix comme principale motivation de visite, les études montrent qu'il ne constitue pas le facteur le plus déterminant de la rétention (Masyhuri, 2022 ; Nguyen & Chanut, 2018). Les clients fortement sensibles aux promotions présentent d'ailleurs un risque de défection plus élevé, leur relation avec l'entreprise étant davantage transactionnelle que fidélisante (Al-Ayed, 2022). Plus fondamentalement, la perception du prix dépend moins de son niveau absolu que du rapport entre le montant payé et la valeur perçue, ce qui explique qu'une défaillance sur la livraison ou la qualité puisse conduire à une impression de surcoût, même en l'absence de variation de prix (Al-Ayed, 2022). La valeur perçue de la prestation conditionne ainsi la légitimité du prix aux yeux du client.

Ces quatre dimensions ont pour point commun de ne pas être directement observables dans les données transactionnelles, mais de s'exprimer explicitement à travers les avis laissés par les clients après leur expérience d'achat.

2.2 Les données pour la prédiction du churn

2.2.1 Données structurées CRM et leurs limites

La prédiction de l'attrition repose en premier lieu sur les données structurées issues des systèmes de Gestion de la Relation Client (Customer Relationship Management, CRM). Cette approche stratégique et organisationnelle vise à gérer l'ensemble du cycle de vie du client en conservant une trace numérique des interactions entre l'entreprise et ses clients, depuis la phase d'acquisition jusqu'à la rétention (Gupta & Zeithaml, 2006). Parmi les variables les plus couramment mobilisées figure la méthode RFM (récence, fréquence et montant), qui permet d'analyser et de segmenter le comportement client à partir des transactions passées (Fader & Hardie, 2009 ; Manzoor et al., 2024).

Ces données CRM présentent plusieurs avantages. Exhaustives et directement disponibles, elles couvrent l'ensemble de la base client sans les contraintes de coût ou d'échantillonnage des enquêtes de satisfaction (Gupta & Zeithaml, 2006), et leur structure facilite leur intégration dans des modèles

prédictifs sans prétraitement complexe. Elles restent particulièrement efficaces pour capturer des comportements observables et mesurables, constituant ainsi une base solide pour la prédiction du churn (Imani et al., 2025).

Néanmoins, ces données demeurent essentiellement descriptives (Fader & Hardie, 2009). Elles indiquent ce que le client a fait sans expliquer les raisons sous-jacentes à ses comportements, et ne fournissent ni éléments psychologiques ni informations contextuelles permettant de comprendre ses motivations (Gupta & Zeithaml, 2006). Elles peuvent ainsi être considérées comme des indicateurs retardés (*lagging indicators*), dans la mesure où elles enregistrent un désengagement déjà entamé, là où les données textuelles permettent potentiellement d'anticiper un départ avant sa manifestation comportementale (Anderson et al., 2004 ; Morgan, 2012).

C'est précisément cette limite qui motive l'exploration de sources d'information complémentaires.

2.2.2 Données textuelles et voix du client

Les données structurées ne représentent qu'une partie de l'information disponible en entreprise. La majorité de l'information produite est en réalité non structurée (Gandomi & Haider, 2015) et correspond principalement à des textes libres et des messages bruts, comme les e-mails, les forums en ligne ou les commentaires clients (De Caigny et al., 2020).

Pour capter la voix des consommateurs, les entreprises exploitent ces informations à travers deux formes de feedback. Le feedback public regroupe le contenu généré par les utilisateurs, incluant le bouche-à-oreille électronique, les avis et les critiques publiés sur les plateformes d'e-commerce ou les réseaux sociaux (Balducci & Marinova, 2018). Le feedback privé correspond aux échanges directs et confidentiels entre le client et l'entreprise, comme les communications individuelles ou les enquêtes de satisfaction (Gupta & Zeithaml, 2006 ; Balducci & Marinova, 2018).

Au sein de ces textes coexistent généralement deux types d'informations. Le contenu attitudinal, de nature subjective, reflète les émotions et opinions du client, tandis que le contenu expérientiel renvoie à des éléments plus descriptifs liés à l'expérience du produit ou du service (Pang & Lee, 2008 ; Archak et al., 2011).

Ensemble, ces informations forment la Voix du Client (Voice of the Customer, VoC) et permettent de dépasser le simple constat de satisfaction ou d'insatisfaction en apportant des éléments explicatifs sur l'expérience vécue par le client (Coussement & Van den Poel, 2008 ; Pang & Lee, 2008 ; Tirunillai & Tellis, 2014). Plusieurs travaux montrent par ailleurs que leur intégration en complément des données CRM améliore significativement les performances des modèles de prédiction du churn (De Caigny et al., 2020), ce qui conduit à les considérer comme des indicateurs avancés (*leading indicators*) par rapport aux données transactionnelles (Anderson et al., 2004 ; Morgan, 2012).

Ces données textuelles ne produisent toutefois un signal prédictif exploitable qu'à condition d'être correctement représentées et intégrées dans un cadre algorithmique, ce qui conduit à exami-

ner les modèles capables de les exploiter.

2.3 Modèles prédictifs pour la classification du churn

2.3.1 Cadre général du machine learning supervisé

L'apprentissage supervisé repose sur un principe simple qui consiste à apprendre à partir d'un ensemble d'observations étiquetées. Un algorithme estime la relation entre des variables d'entrée X et une variable cible Y afin de généraliser cette relation à de nouvelles observations (James et al., 2023). L'objectif principal de cet algorithme est de produire des prédictions précises sur des données non observées, mais également de comprendre la structure des relations entre les variables cibles et les variables prédictives (James et al., 2023).

Selon la nature de la variable cible, on distingue deux grandes familles de problèmes. Les problèmes de régression concernent une variable continue, telle qu'une valeur réelle ou un scalaire, tandis que les problèmes de classification portent sur une variable catégorielle, comme la prédiction du départ d'un client (Murphy, 2012).

Les algorithmes d'apprentissage supervisé peuvent être regroupés en plusieurs grandes familles selon leur principe de modélisation. Les modèles linéaires, comme la régression linéaire ou encore logistique, établissent une relation directe entre les variables d'entrée et la variable cible (Murphy, 2012). Leur principal atout réside dans leur interprétabilité, chaque variable se voyant attribuer un coefficient traduisant sa contribution marginale à la prédiction (Manzoor et al., 2024). Les arbres de décision quant à eux partitionnent l'espace des données à travers une succession de règles binaires, offrant une grande lisibilité mais présentant une instabilité importante face aux variations de l'échantillon d'entraînement (James et al., 2023 ; Manzoor et al., 2024). Les méthodes d'ensemble, qui agrègent les prédictions de plusieurs arbres, permettent de réduire cette instabilité et figurent parmi les approches les plus performantes sur des données tabulaires complexes, au prix d'une perte d'interprétabilité (James et al., 2023 ; Manzoor et al., 2024). Enfin, le deep learning repose sur des architectures de réseaux de neurones multicouches particulièrement adaptées aux données non structurées telles que le texte ou les images, mais nécessitant des volumes de données importants et une forte puissance de calcul pour être efficaces (Manzoor et al., 2024 ; Imani et al., 2025).

Dans ce cadre, la prédiction du churn se distingue d'un problème de classification classique par deux difficultés structurelles. La première difficulté réside dans le déséquilibre des classes. Les churners étant minoritaires, les modèles standards tendent à favoriser la classe majoritaire au détriment des cas d'intérêt (Manzoor et al., 2024). La seconde concerne l'asymétrie du coût des erreurs, une non détection entraînant la perte du client et de ses revenus futurs, tandis qu'une action de rétention sur un client non churner reste relativement peu coûteuse (Verbeke et al., 2011 ; Jahromi et al., 2014).

2.3.2 Modèles pour des données tabulaires déséquilibrées

Dans la continuité des défis propres à la prédiction du churn, la régression logistique constitue une première approche de référence.

Elle estime la probabilité d'appartenance à une classe en transformant une combinaison linéaire des variables d'entrée en une valeur comprise entre 0 et 1 grâce à la fonction sigmoïde (Verbeke et al., 2011). Chaque variable se voit attribuer un coefficient qui traduit sa contribution marginale à la probabilité de la classe cible, ce qui permet d'interpréter directement l'effet des différents prédicteurs et constitue un atout majeur dans les contextes où l'explicabilité est essentielle (Manzoor et al., 2024). Bien que relativement simple, la régression logistique demeure compétitive face à des approches plus complexes lorsque les données sont correctement structurées et constitue à ce titre un point de référence dans les études comparatives sur la prédiction du churn (Verbeke et al., 2011 ; Manzoor et al., 2024).

Les approches ensemblistes, et plus particulièrement le XGBoost, offrent une alternative plus flexible et plus performante que la régression logistique. Plutôt que d'estimer une relation globale entre les variables et la cible, il construit de manière séquentielle une série d'arbres de décision, chacun entraîné à corriger les erreurs du précédent (Manzoor et al., 2024). Le taux d'apprentissage contrôle l'ampleur de ces corrections et permet une convergence progressive tout en limitant le risque de surapprentissage. Cette logique itérative lui confère un avantage structurel face au déséquilibre des classes, les observations mal classées voyant leur poids augmenter à chaque itération, ce qui oriente progressivement le modèle vers les cas les plus difficiles à détecter, dont font typiquement partie les churners (Jahromi et al., 2014 ; Imani et al., 2025). Sa supériorité empirique sur les données tabulaires déséquilibrées est largement documentée dans la littérature, où il surpasse généralement des modèles plus simples (Manzoor et al., 2024 ; Imani et al., 2025).

Par ailleurs, les méthodes de boosting présentent un avantage décisif sur les architectures de deep learning, qui nécessitent de très larges volumes de données pour être performantes et dont la complexité computationnelle n'est pas justifiée sur des datasets de taille modérée (Manzoor et al., 2024 ; Imani et al., 2025).

Dans ce contexte, ces modèles restent particulièrement adaptés aux données structurées, mais leur performance dépend directement de la qualité et de la richesse des variables disponibles. Lorsque les données transactionnelles ne suffisent pas à capturer l'ensemble du signal de churn, la mobilisation des données textuelles comme source d'information complémentaire devient centrale.

2.4 Représentation du texte pour l'analyse prédictive

2.4.1 Approches creuses et les embeddings statiques

Les représentations de texte visent à transformer le langage naturel en objets mathématiques exploitables par des algorithmes. Les premières approches dites creuses reposent sur des vecteurs de grande dimension majoritairement nuls. Le modèle Bag-of-Words représente un texte comme un simple histogramme de mots en ignorant leur ordre, ce qui implique par exemple que des termes comme « livraison » et « retard » ne sont pas plus proches que « livraison » et « prix » (Blei et al., 2003). Le TF-IDF améliore cette représentation en pondérant les termes selon leur fréquence locale et leur rareté globale, ce qui permet de mettre en avant les mots discriminants (Manning et al., 2008). Ces approches restent utiles pour la recherche d'information mais ne capturent pas les relations sémantiques entre les mots ni leur contexte d'apparition.

Les embeddings statiques ont ensuite permis de dépasser ces limites en représentant chaque mot par un vecteur dense dans un espace continu, fondé sur l'hypothèse distributionnelle selon laquelle des mots apparaissant dans des contextes similaires possèdent des significations proches (Mikolov et al., 2013). Word2Vec apprend ces représentations à partir du contexte local tandis que GloVe exploite des co-occurrences globales.

Ces méthodes réduisent fortement la dimensionnalité et introduisent une proximité sémantique entre les mots, mais elles restent limitées par leur caractère statique, un même mot conservant une représentation identique quel que soit son contexte (Mao et al., 2024). Ces approches sont par ailleurs sensibles au domain shift, qui correspond à la baisse de performance d'un modèle lorsqu'il est appliqué à un domaine différent de celui de son entraînement (Taboada et al., 2011).

2.4.2 Embeddings contextuels : Transformers, BERT et SentenceTransformer

L'architecture Transformer, introduite en 2017, marque une rupture importante avec les approches précédentes en abandonnant le traitement séquentiel au profit d'un traitement simultané des mots (Reimers & Gurevych, 2019). Son fonctionnement repose sur le mécanisme d'auto-attention, qui permet au modèle d'analyser l'ensemble d'une séquence en tenant compte des relations entre les mots, quelle que soit leur distance dans le texte (Reimers & Gurevych, 2019). Les représentations produites deviennent ainsi contextuelles, un même mot pouvant recevoir une signification différente selon la phrase dans laquelle il apparaît.

Cette architecture a ensuite donné naissance à deux grandes familles de modèles. Les modèles de type encodeur, tels que BERT, produisent des représentations contextuelles particulièrement adaptées aux tâches de classification après fine-tuning (Devlin et al., 2019). SentenceTransformer s'inscrit dans cette logique en étant optimisé pour générer des représentations denses de phrases entières et mesurer leur similarité sémantique (Reimers & Gurevych, 2019). À l'inverse, les mo-

dèles de type décodeur, comme GPT ou LLaMA, appartiennent à la famille des grands modèles de langage (Large Language Models, LLM). Reposant sur une logique générative, ils sont capables de réaliser une grande variété de tâches à partir de simples instructions textuelles appelées prompts (Zhao et al., 2023).

Malgré leur flexibilité, ces modèles restent limités par leur mémoire paramétrique, c'est-à-dire par les connaissances acquises lors du pré-entraînement. Celles-ci sont figées à la date d'entraînement du modèle et leur mise à jour nécessite des phases coûteuses de réentraînement ou de fine-tuning (Zhao et al., 2023 ; Klesel & Wittmann, 2025).

2.4.3 LLM génératifs et architecture RAG

Pour pallier les limites d'une mémoire figée, les travaux récents se sont orientés vers l'intégration d'une mémoire non paramétrique. Celle-ci correspond à des informations stockées explicitement en dehors du modèle, généralement sous la forme d'une base documentaire ou d'un index vectoriel dense (Lewis et al., 2020 ; Klesel & Wittmann, 2025).

Dans cette logique, les architectures hybrides de type Retrieval-Augmented Generation (RAG) combinent mémoire paramétrique et mémoire non paramétrique afin de guider la génération du modèle de langage à partir d'informations récupérées dynamiquement (Lewis et al., 2020 ; Klesel & Wittmann, 2025).

Le fonctionnement d'un pipeline RAG repose d'abord sur une phase d'indexation, réalisée en amont de toute requête utilisateur. Les documents issus de la base de connaissances sont nettoyés puis découpés en segments de texte appelés chunks. Ces segments sont ensuite transformés en vecteurs grâce à un modèle d'embedding avant d'être stockés dans une base vectorielle constituant la mémoire non paramétrique (Asai et al., 2023).

Une fois cette mémoire externe construite, le pipeline RAG proprement dit débute par une phase de récupération (*retrieval*). Lorsqu'une requête est formulée, celle-ci est convertie en vecteur à l'aide du même modèle d'embedding que celui utilisé lors de l'indexation. Le système de récupération calcule alors la similarité entre le vecteur de la requête et ceux des différents chunks afin d'identifier les passages les plus pertinents (Asai et al., 2023). Ces documents sont ensuite injectés dans le prompt lors des étapes d'augmentation et de génération. Le modèle de langage s'appuie simultanément sur ce contexte externe et sur ses connaissances paramétriques pour produire une réponse plus contextualisée et plus factuelle (Lewis et al., 2020 ; Asai et al., 2023).

Le principal intérêt du RAG réside dans sa capacité à actualiser les connaissances mobilisées par le modèle sans nécessiter de réentraînement complet. Les informations utilisées peuvent ainsi être mises à jour directement via les documents présents dans l'index externe, réduisant fortement les coûts et la complexité associés aux modèles de langage (Lewis et al., 2020).

Cependant, l'efficacité du pipeline reste étroitement liée à la qualité du mécanisme de retrieval et à la fiabilité de la base documentaire. Si la base contient des informations contrefactuelles ou peu pertinentes, celles-ci peuvent orienter la génération vers des conclusions inexactes (Klesel &

Wittmann, 2025 ; Asai et al., 2023). Des variantes plus avancées ont depuis été proposées, notamment l'Advanced RAG et le Self-RAG, qui introduisent des mécanismes supplémentaires comme le re-ranking des passages récupérés ou l'auto-évaluation de la pertinence des sources (Gao et al., 2024 ; Asai et al., 2023).

2.4.4 Positionnement du mémoire et contribution scientifique

Les architectures RAG sont principalement utilisées dans des tâches liées à l'accès à l'information et à la génération de texte. Leur application la plus courante concerne les systèmes de questions-réponses, couvrant aussi bien les questions ouvertes nécessitant plusieurs étapes de raisonnement que les réponses longues ou les questions à choix multiples (Gao et al., 2023).

L'analyse du churn mobilise les données textuelles selon une logique différente. Les textes sont généralement transformés en variables exploitables par des modèles prédictifs, que ce soit via des méthodes de NLP classiques, des représentations vectorielles ou des dictionnaires de sentiments prédéfinis (De Caigny et al., 2020 ; Chajia & Nfaoui, 2024). Dans ces approches, les modèles de langage sont donc essentiellement utilisés pour encoder le texte, et non pour interagir avec une base documentaire via une architecture RAG.

En revanche, la capacité d'une architecture comme le RAG à produire des variables exploitables dans un cadre analytique ou prédictif reste très peu étudiée. À ce stade, il existe un manque d'études comparant directement cette architecture aux approches textuelles classiques, afin d'évaluer leur capacité à générer des caractéristiques pertinentes (*features*) destinées à alimenter un modèle de scoring ou de prédiction. Ce mémoire propose de combler ce gap en comparant empiriquement quatre approches de traitement du signal client, afin d'évaluer dans quelle mesure une extraction sémantique structurée via RAG permet d'améliorer la prédiction du churn tout en produisant des features interprétables et actionnables dans un contexte métier.

Méthodologie

Ce chapitre présente le dispositif méthodologique mis en place dans le but d'évaluer l'apport de différentes représentations textuelles dans la prédiction du churn client. L'étude repose sur une démarche expérimentale comparative, dans laquelle plusieurs approches sont évaluées dans un cadre contrôlé afin d'isoler l'effet du type de représentation textuelle mobilisée.

Après la présentation du dataset et de la construction de la variable cible, les différentes conditions expérimentales, les modèles utilisés ainsi que les protocoles d'évaluation sont détaillés.

3.1 Positionnement épistémologique et design de recherche

L'objectif de ce mémoire est d'évaluer dans quelle mesure différentes formes de représentation du texte contribuent à l'amélioration de la prédiction du churn client. Dans cette perspective, trois axes d'analyse sont retenus. Le premier consiste à mesurer l'apport des données textuelles par rapport à une approche baseline reposant uniquement sur des variables structurées. Le second vise à comparer deux approches supervisées de représentation textuelle, des *embeddings* contextuels et une extraction sémantique structurée via un *pipeline RAG*. L'objectif est double, puisqu'il s'agit à la fois d'évaluer leur performance prédictive ainsi que d'examiner leur degré d'interprétabilité. Enfin, une troisième approche, fondée sur un modèle de langage zero-shot, est introduite afin d'évaluer la capacité d'un LLM à inférer directement le churn à partir du texte, constituant ainsi une référence non supervisée.

Pour répondre à ces objectifs, ce mémoire adopte une démarche empirique reposant sur l'expérimentation. Il s'agit de comparer, dans un cadre contrôlé, les performances relatives de plusieurs approches afin de combler un manque identifié dans la littérature. Le dispositif expérimental repose sur un principe central dans lequel toutes les conditions sont strictement comparables, à l'exception d'un seul facteur, à savoir le type de représentation textuelle mobilisé. Ainsi, la population étudiée, les modèles utilisés, les métriques d'évaluation ainsi que le découpage des données sont maintenus constants entre les différentes configurations.

Ce contrôle strict des variables permet d'isoler l'effet propre de la représentation du texte sur la performance prédictive. Dans cette logique, quatre conditions expérimentales sont définies, correspondant à quatre niveaux d'intégration du texte allant de leur absence complète à leur exploitation directe par un modèle de langage, comme synthétisé dans le Tableau 1.

TABLE 1 – Synthèse des quatre conditions expérimentales

Condition	Rôle	Features	Modèles	Train
M1	Baseline structurée	Données structurées seules (18 features)	LR + XGBoost	27 213
M2	Approche comparée	Données structurées + embeddings (402 features)	LR + XGBoost	27 213
M3	Approche comparée	Données structurées + RAG (26 features)	LR + XGBoost	1 500
M4	Référence non supervisée	Reviews brutes	GPT-4o mini	—

Le design repose sur deux points de comparaison. M1 constitue une baseline structurée correspondant au niveau atteignable sans information textuelle, tandis que M4 sert de référence non supervisée évaluant les performances d'un LLM généraliste sans apprentissage spécifique ni ingénierie de features. À ces deux points de référence s'ajoutent deux approches supervisées au cœur de l'analyse, M2 basé sur des embeddings contextuels et M3 reposant sur un pipeline RAG. L'analyse porte principalement sur la confrontation entre ces deux dernières.

3.2 Données

3.2.1 Présentation du *dataset*

La base de données mobilisée dans ce mémoire est le Brazilian E-Commerce Public Dataset publié par Olist sur la plateforme Kaggle. À notre connaissance, ce dataset fait partie des rares jeux de données publics qui combine, à l'échelle transactionnelle, des données structurées réelles et complètes ainsi que des avis textuels directement associés à chaque commande. Cette double nature en fait un support particulièrement adapté à cette analyse comparative. Plus largement, les jeux de données publics combinant simultanément données structurées de type CRM, avis textuels et historique demeurent rares.

Cette contrainte, inhérente aux données disponibles dans le domaine, justifie le recours à un proxy d'insatisfaction comme variable cible.

Le *dataset* couvre une période allant de septembre 2016 à août 2018 et a fait l'objet d'un travail de préparation afin de garantir la cohérence de l'analyse. Dans un premier temps, seules les commandes effectivement livrées ont été conservées. Le dataset est ainsi passé de 99 441 à 96 478 observations. Les commandes ne disposant pas de note de satisfaction ont ensuite été supprimées, aboutissant à un échantillon final de 95 832 clients. Chaque observation correspond à un client unique associé à une seule commande, ce qui représente une contrainte structurelle forte du *dataset*.

Enfin, 41 % des clients ont laissé un commentaire textuel, soit 38 876 observations, rédigées en portugais. Cette caractéristique délimite la population sur laquelle l'ensemble des conditions expérimentales seront évaluées, comme détaillé ultérieurement.

3.2.2 Construction de la variable cible

La structure du *dataset* impose une contrainte méthodologique centrale, dans la mesure où chaque client n'est associé qu'à une seule commande, rendant impossible la mesure du churn comportemental au sens classique, fondé sur l'observation du non-retour. Cette limitation rend nécessaire le recours à un *proxy* attitudinal, reposant sur un signal d'insatisfaction exprimé.

Ainsi, la variable cible est construite à partir de la note de satisfaction attribuée par le client. Un client est considéré comme *churner* lorsque son *score de review* est inférieur ou égal à 2, et non churner dans le cas contraire. Ce seuil est justifié par la distribution bimodale des notes, caractérisée par des pics marqués aux extrêmes, où les notes 1 et 2 traduisent une insatisfaction explicite (Voir Annexe 1).

Bien que ce choix ne corresponde pas à une mesure directe du churn comportemental, il repose sur l'idée que les niveaux d'insatisfaction les plus marqués concentrent l'essentiel de l'information prédictive. Cette intuition est cohérente avec des travaux montrant que la relation entre satisfaction et fidélité n'est pas linéaire, mais asymétrique. En particulier, seules les évaluations les plus faibles ont un impact significatif sur le comportement futur du client, tandis que les niveaux de satisfaction intermédiaires ou élevés influencent peu la fidélité (van Doorn & Verhoef, 2008).

Il est toutefois important de préciser que ce proxy ne reflète qu'une partie du phénomène de churn. Certains clients peuvent quitter la plateforme sans jamais exprimer leur insatisfaction à travers un commentaire négatif. Leur désengagement reste alors invisible dans les données disponibles. À l'inverse, une mauvaise évaluation ponctuelle ne traduit pas nécessairement un départ définitif. Certains clients insatisfaits à un moment donné peuvent continuer à effectuer des achats par la suite. De plus, un client globalement satisfait peut malgré tout se tourner vers un concurrent en raison d'une offre plus avantageuse, sans que cela apparaisse dans les signaux observés. Ces situations montrent qu'il existe un écart entre l'insatisfaction exprimée et le churn réellement observable dans les comportements des clients.

Dans cette perspective, l'objectif n'est pas de reconstituer une mesure parfaite du churn, mais d'identifier un seuil de rupture traduisant une insatisfaction forte, susceptible d'être associée à un risque de désengagement. Ce choix constitue donc un compromis méthodologique lié à l'absence de données longitudinales, tout en restant cohérent avec les mécanismes comportementaux mis en évidence dans la littérature.

Sur cette base, le taux de churn s'élève à 12,8 % sur l'ensemble du *dataset*, contre 25,2 % sur le sous-ensemble des clients ayant laissé un commentaire, qui constitue la population effective de l'étude.

3.2.3 Construction du dataset commun et split temporel

La définition de la variable cible étant établie, la construction du périmètre d'analyse peut être précisée. Afin de garantir la comparabilité des conditions expérimentales, l'ensemble des analyses est conduit sur un *dataset* commun restreint aux clients ayant laissé un commentaire textuel, soit 38 876 observations (40,6 % de la population totale). Ce choix émane d'une exigence méthodologique visant à garantir que l'ensemble des conditions, de M1 à M4, soit évalué sur une population strictement identique. Sans cette restriction, comparer une condition sur l'ensemble des 95 832 clients à une autre construite sur un sous-échantillon plus restreint introduirait un biais de sélection rendant toute comparaison invalide.

Ce choix entraîne toutefois une limite importante. Les clients qui expriment un avis écrit constituent une population spécifique, potentiellement plus engagée ou plus polarisée dans leur expérience. Cette différence se reflète empiriquement dans la distribution de la variable cible, avec un taux de churn d'environ 25 % sur le sous-échantillon textuel, contre 12,8 % sur l'ensemble du *dataset*. Par conséquent, les performances observées, en particulier pour la condition M1, ne sont pas directement généralisables à la population complète. Ce biais est assumé, car il constitue la contrepartie nécessaire d'un cadre expérimental contrôlé.

Ainsi, un découpage temporel des données est mis en place afin de préserver la validité du protocole d'évaluation. Un *split* aléatoire introduirait un risque de fuite de données, en permettant au modèle d'exploiter implicitement des observations futures lors de l'entraînement. À l'inverse, le découpage chronologique reproduit les conditions réelles de déploiement, dans lesquelles un modèle apprend à partir des données passées afin de prédire des comportements futurs.

Le *dataset* est ainsi divisé en trois sous-ensembles construits sur cette même population : un jeu d'entraînement (70 %, soit 27 213 observations de septembre 2016 à avril 2018), un jeu de validation (15 %, soit 5 831 observations, d'avril à juin 2018) et un jeu de test (15 %, soit 5 832 observations, de juin à août 2018). Ce dernier est figé avant toute expérimentation et n'est mobilisé qu'une seule fois pour l'évaluation finale, afin de préserver une estimation non biaisée des performances.

Ce découpage met en évidence un *drift* temporel du churn, avec un taux décroissant entre le jeu d'entraînement (25,2 %), la validation (21,7 %) et le test (19,2 %). Deux mécanismes peuvent expliquer cette évolution. D'une part, un biais de maturité, dans lequel les clients les plus récents disposent de moins de recul pour exprimer une insatisfaction. D'autre part, un effet de sélection temporelle, les premiers utilisateurs de la plateforme étant plus enclins à laisser des avis, notamment négatifs. Ce déséquilibre est conservé en l'état, la cohérence chronologique étant privilégiée à une stratification artificielle qui romprait la logique temporelle du dispositif.

Enfin, l'ensemble des étapes de prétraitement est strictement contraint au jeu d'entraînement pour éviter toute fuite d'information. Le *target encoding* ainsi que les paramètres de normalisation sont estimés exclusivement sur ce dernier, puis appliqués à l'identique aux jeux de validation et de test, sans recalcul.

3.3 Conditions expérimentales

3.3.1 Condition M1 : Données structurées seules

La première condition M1 repose exclusivement sur les données structurées et constitue la *baseline* de l'étude. Elle permet d'établir un niveau de performance de référence en l'absence de toute information textuelle, en s'appuyant uniquement sur des variables transactionnelles.

L'ensemble des 18 *features* mobilisées est construit selon une logique fonctionnelle. Il inclut des variables de type RFM (*recency_days, frequency, monetary*), des variables comportementales liées à la transaction (*n_items, n_sellers, delivery_delay_days, payment_value*), des variables de paiement encodées en one-hot, ainsi qu'une variable produit (*product_weight_g*). La variable *frequency*, bien que constante dans le *dataset* et égale à 1 pour tous les clients, est conservée afin de maintenir la cohérence avec les approches classiques de segmentation client (Voir Annexe 2).

Deux variables catégorielles à forte cardinalité, *state* et *category*, sont encodées via un *target encoding*, calculé exclusivement sur le jeu d'entraînement, puis appliqué à l'identique aux jeux de validation et de test, afin d'éviter toute fuite d'information. De manière cohérente, l'ensemble des transformations de prétraitement est également contraint au jeu d'entraînement. La normalisation est réalisée via un StandardScaler ajusté uniquement sur ce dernier, et les valeurs manquantes sont imputées par la médiane observée sur ce même jeu.

Dans cette configuration, M1 est entraîné sur l'intégralité du jeu d'entraînement, soit 27 213 observations.

3.3.2 Condition M2 : Données structurées + *embeddings* contextuels

La condition M2 étend la *baseline* en intégrant l'information textuelle issue des avis clients sous forme d'*embeddings*. L'objectif est d'évaluer dans quelle mesure l'ajout d'un signal sémantique non structuré améliore la performance prédictive par rapport aux données transactionnelles seules.

Les reviews sont encodés à l'aide du modèle *paraphrase-multilingual-MiniLM-L12-v2* (*Sentence-Transformers*), choisi pour sa capacité à capturer des similarités sémantiques fines, tout en étant adapté au portugais. Chacune d'entre elles, obtenue par concaténation du titre et du message, est projetée dans un espace vectoriel dense de dimension 384. Cette représentation permet de transformer le contenu textuel hétérogène en variables quantitatives directement exploitables par le modèle.

Les *embeddings* ainsi obtenus sont concaténés aux 18 variables structurées de M1, conduisant à un vecteur final de 402 *features*. Afin de préserver l'équilibre entre les deux sources d'information, une normalisation est appliquée séparément sur les variables structurées et sur les *embeddings* avant concaténation, de manière à éviter qu'un bloc ne domine l'autre en raison d'un effet d'échelle.

Le choix d'entraîner M2 sur l'ensemble du jeu d'entraînement repose sur une analyse empirique. Une *learning curve*, estimée sur des tailles d'échantillon croissantes, allant de 500 à 27 000 observations, met en évidence un plateau de performance en PR-AUC à partir d'environ 20 000

observations (voir Annexe 4). L'utilisation de l'intégralité des données disponibles pour l'entraînement est ainsi retenue, la performance se stabilisant au-delà de ce seuil.

Toutefois, afin de permettre une comparaison à taille d'entraînement fixe avec la condition M3, M2 est également entraîné sur un sous-échantillon de 1 500 observations, identique à celui utilisé pour M3. Cette analyse complémentaire vise à isoler l'effet propre de la représentation textuelle, en supprimant l'avantage de volume dont bénéficie M2 dans l'analyse principale.

3.3.3 Condition M3 : Données structurées + pipeline *RAG*

Architecture du pipeline

La condition M3 introduit un *pipeline* de type *Retrieval-Augmented Generation* (*RAG*), dont l'objectif n'est pas uniquement d'exploiter le texte, mais de le structurer à partir d'un cadre théorique explicite. Contrairement à M2 qui repose sur une représentation non interprétable, M3 vise à produire des variables sémantiques directement alignées avec les déterminants du churn identifiés dans la littérature.

La base de connaissances, construite en Annexe 3, est issue des articles mobilisés dans la revue de littérature. Elle se compose de 41 passages organisés selon quatre dimensions clés du churn en e-commerce, à savoir la livraison, la qualité du produit, le service client et le prix. Chaque passage est explicitement associé à sa dimension, ce qui permet d'ancrer le processus de *retrieval* dans une structure conceptuelle définie en amont.

La sélection de ces passages a été réalisée manuellement à partir de ces mêmes travaux. La lecture de Masyhuri (2022), Al-Ayed (2022), Nguyen & Chanut (2018), De Caigny et al. (2020), Tirunillai & Tellis (2014) et Archak et al. (2011) a permis d'identifier les quatre dimensions les plus fréquemment mobilisées comme déterminants du churn ou d'insatisfaction en e-commerce. Pour chaque dimension, les passages les plus représentatifs et les plus explicites ont été retenus, c'est-à-dire ceux exprimant le plus clairement le lien entre la dimension concernée et l'insatisfaction ou le désengagement client.

Ces passages sont encodés à l'aide du modèle paraphrase-multilingual-MiniLM-L12-v2 (Sentence-Transformers), le même modèle multilingue que celui utilisé pour l'encodage des reviews dans le mécanisme de retrieval. Ce choix garantit que les représentations vectorielles des passages et des reviews sont produites dans un même espace sémantique.

Le mécanisme de *retrieval* repose quant à lui sur *FAISS*, utilisé pour calculer la similarité cosinus entre les *embeddings* des *reviews* et ceux des passages de la base de connaissances. Pour chaque observation, les k passages les plus proches sémantiquement sont sélectionnés. Afin de ne pas fixer ce paramètre de manière arbitraire, une analyse de sensibilité est conduite en testant plusieurs valeurs de k pour évaluer la stabilité des corrélations entre les scores RAG produits et la note de satisfaction (voir Annexe 5).

Avant leur transmission au pipeline, les reviews sont préparées par concaténation du titre et du message, sans prétraitement linguistique supplémentaire. Dans un pipeline RAG reposant sur un

LLM, le texte brut constitue une source d'information plus riche qu'un texte appauvri. Les nuances stylistiques, les marqueurs émotionnels et les tournures idiomatiques contribuent au contexte sémantique que GPT-4o mini exploite pour produire ses scores. Réduire cette richesse reviendrait à priver le modèle d'une partie du signal disponible.

Toutefois, la review et chaque passage récupéré par FAISS sont tronqués à 200 caractères dans le prompt afin de maîtriser la taille totale du contexte envoyé au modèle. Pour les reviews, cette contrainte n'a presque pas d'impact, dans la mesure où 98,4 % d'entre elles font moins de 200 caractères, avec une longueur médiane de 52 caractères. Pour les passages récupérés, la troncature affecte uniquement la fin de la dernière phrase, sans couper d'information sémantiquement nouvelle. Elle n'affecte donc pas le contenu informatif dans la majorité des cas.

Ces passages sont ensuite intégrés au *prompt*, conjointement à la *review*, afin de construire un contexte enrichi. Ce dernier est transmis à *GPT-4o mini*, qui génère quatre scores compris entre 0,0 et 1,0, chacun correspondant à une des dimensions identifiées. Le paramètre de température est fixé à 0 afin d'assurer la reproductibilité des sorties.

Lorsque certaines dimensions ne sont pas explicitement mentionnées dans le texte, une valeur neutre de 0,5 est attribuée. Cette convention introduit toutefois une ambiguïté entre absence d'information et neutralité réelle. Pour y remédier, quatre variables binaires supplémentaires sont introduites, indiquant explicitement si chaque dimension est mentionnée dans la *review* ou non.

Au terme de ce processus, le *pipeline RAG* produit un ensemble de 8 variables textuelles, dont 4 scores continus et 4 indicateurs binaires de mention, combinées aux 18 variables structurées pour former un vecteur final de 26 *features*. L'ensemble de ce processus est synthétisé dans la Figure 1 ci-dessous.

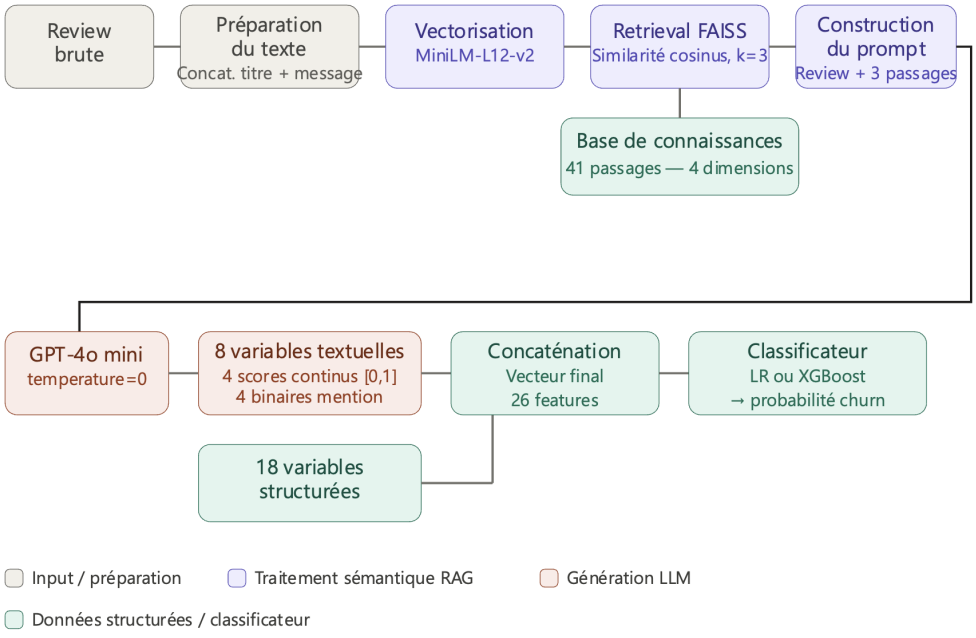


FIGURE 1 – Architecture du pipeline RAG

Le recours à *GPT-4o mini* s'explique par un compromis entre performance, coût et robustesse.

Des alternatives locales ont été testées à cet effet, notamment *Phi-3 mini* et *Mistral 7B* via *Ollama*, mais se sont révélées instables sur le corpus portugais, avec des sorties peu fiables et des hallucinations fréquentes.

Détermination de la taille d'échantillon d'entraînement

Le recours au *pipeline RAG* implique un coût computationnel significatif lié aux appels API, rendant l'entraînement sur l'ensemble du jeu d'entraînement peu pertinent a priori. La taille de l'échantillon retenue repose alors sur une validation empirique plutôt que d'un choix arbitraire.

Une *learning curve* préliminaire est estimée en évaluant les performances pour des tailles d'échantillon croissantes (500, 750, 1 000, 1 500 et 2 000 observations), avec des hyperparamètres fixés par défaut. Elle vise uniquement à identifier le point de saturation de la courbe d'apprentissage, indépendamment du paramétrage optimal (Voir Annexe 6). La taille finalement retenue à l'issue de cette analyse est présentée dans la section Résultats et Analyses.

Validation de cohérence des scores *RAG*

Avant leur intégration dans les modèles prédictifs, les scores issus du *pipeline RAG* font l'objet d'une validation empirique afin de vérifier qu'ils capturent effectivement un signal sémantique pertinent, et non un artefact arbitraire du modèle génératif.

Cette validation est conduite sur les 1 500 observations du train M3. Les résultats mettent en évidence des corrélations significatives entre les scores *RAG* et la note de satisfaction. Les dimensions livraison ($r = 0{,}646$, $p < 0{,}001$) et qualité produit ($r = 0{,}514$, $p < 0{,}001$) présentent les corrélations les plus fortes, suivies du service client ($r = 0{,}419$, $p < 0{,}001$), tandis que la dimension prix reste plus faiblement corrélée ($r = 0{,}167$, $p < 0{,}001$). Ces écarts sont alignés avec la nature des avis post-achat, dans lesquels les problèmes liés à la livraison et à la qualité du produit sont plus fréquemment exprimés que les considérations de prix.

Cette relation de corrélation est renforcée par l'analyse des scores moyens par niveau de satisfaction, qui met en évidence une progression globalement monotone (voir Annexe 7). À mesure que la note augmente, les scores associés aux différentes dimensions augmentent également. Ce schéma confirme que les variables générées reflètent de manière structurée l'expérience client. De légères irrégularités locales peuvent apparaître, notamment entre certaines notes adjacentes, mais ne remettent pas en cause la tendance globale.

Par ailleurs, les distributions des scores révèlent des disparités importantes entre dimensions. Le service client et le prix sont rarement mentionnés explicitement, dans respectivement 12,1 % et 3,5 % des reviews. À l'inverse, la livraison et la qualité du produit sont présentes dans 53,4 % et 41,7 % des avis. Ce résultat suggère que l'insatisfaction des clients Olist s'exprime principalement à travers ces deux dernières dimensions, tandis que le prix et le service client font l'objet d'une verbalisation beaucoup plus rare.

3.3.4 Condition M4 : LLM zero-shot

La condition M4 repose sur une approche zero-shot dans laquelle *GPT-4o mini* reçoit uniquement la *review* brute et produit directement une probabilité de churn comprise entre 0,0 et 1,0, sans phase d'entraînement ni construction explicite de variables. Le choix de la probabilité continue est nécessaire afin de permettre l'optimisation du seuil de décision ainsi que le calcul des métriques d'évaluation communes aux autres conditions. Cette configuration constitue une référence non supervisée, permettant d'évaluer ce qu'un *LLM* généraliste peut accomplir sans apprentissage sur des données étiquetées.

En l'absence d'apprentissage supervisé, le jeu de validation est mobilisé dans le seul but de calibrer le seuil de décision. Un sous-échantillon de 500 observations, stratifié sur la variable de churn et sélectionné en fin de période de validation, est utilisé pour déterminer le seuil maximisant le *F1-score*. Le seuil retenu, 0,800, est ensuite appliqué à l'identique au jeu de test, sans ajustement supplémentaire.

Cette étape de calibration du seuil, bien qu'elle repose sur un sous-échantillon étiqueté, ne constitue pas un apprentissage supervisé au sens strict. Elle vise uniquement à déterminer le seuil de décision optimal sur la courbe precision-recall, de manière cohérente avec les conditions supervisées, dont les seuils sont également optimisés sur le jeu de validation. À l'inverse, fixer arbitrairement un seuil à 0,5 introduirait un biais défavorable à M4 et compromettrait l'équité de la comparaison entre les différentes conditions expérimentales.

3.4 Modèles et optimisation des hyperparamètres

Dans le cadre de ce mémoire, deux modèles supervisés, à savoir la régression logistique et le *XGBoost*, sont mobilisés pour les conditions M1 à M3, la condition M4 relevant d'une approche non supervisée. Leur utilisation conjointe répond à une logique de complémentarité entre interprétabilité et performance, les architectures de deep learning ayant été écartées faute d'un volume de données suffisant (Manzoor et al., 2024).

La régression logistique est utilisée comme modèle de référence. En tant que modèle linéaire, elle permet d'estimer la contribution marginale de chaque variable et offre un haut niveau d'interprétabilité (Verbeke et al., 2011 ; Manzoor et al., 2024). Elle constitue ainsi un point de comparaison essentiel pour évaluer si l'apport de modèles plus complexes se traduit effectivement par un gain de performance. Le déséquilibre de classes est pris en compte via le paramètre *class_weight = balanced*, et une régularisation L2 est appliquée via le paramètre *C* afin de limiter le surapprentissage. Le nombre maximal d'itérations est fixé à 1 000 afin de garantir la convergence de l'algorithme d'optimisation.

Le modèle *XGBoost* est mobilisé quant à lui en complément pour sa capacité à capturer des relations non linéaires et des interactions complexes entre variables. Particulièrement adapté aux données tabulaires déséquilibrées (Manzoor et al., 2024 ; Imani et al., 2025), il intègre directe-

ment des mécanismes de régularisation, notamment L1 via *reg_alpha* et L2 via *reg_lambda*. Le déséquilibre de classes est traité via le paramètre *scale_pos_weight*, défini comme le ratio entre *non-churners* et *churners*.

Le traitement par pondération a été privilégié aux approches par rééchantillonnage. SMOTE a été envisagé mais écarté, la génération d'observations synthétiques étant incompatible avec le pipeline RAG, qui repose sur des reviews réelles, ainsi qu'avec la logique de split temporel. L'undersampling a également été exclu en raison de la perte d'information qu'il induirait sur un jeu de données de taille modérée.

Par ailleurs, l'optimisation des hyperparamètres est réalisée par *GridSearch* sur le seul jeu de validation, afin de préserver l'intégrité du jeu de test. La *PR-AUC* est utilisée comme métrique d'optimisation, en cohérence avec le caractère déséquilibré de la variable cible. Une fois les hyperparamètres optimaux identifiés, ceux-ci sont figés et appliqués à l'identique au jeu de test, sans recalcul. Les grilles testées pour chaque condition sont détaillées en Annexe 8.

Enfin, un surapprentissage modéré est observé pour certaines configurations. En condition M1, cet écart atteint 0,15 environ et reflète la limite intrinsèque des *features* transactionnelles face au *drift* temporel. Il ne s'agit pas d'un surapprentissage corrigeable par régularisation. En condition M3, l'écart est plus modéré, compris entre 0,07 et 0,09, et s'explique par la taille réduite de l'échantillon d'entraînement. Dans les deux cas, ce phénomène est interprété comme structurel et documenté comme une limite inhérente au dispositif expérimental.

3.5 Évaluation

La performance des modèles est évaluée principalement à l'aide de la *PR-AUC*, métrique particulièrement adaptée aux contextes de déséquilibres de classes. Cette dernière permet de se concentrer sur la capacité du modèle à identifier correctement la classe minoritaire, les *churners*, en intégrant explicitement le compromis entre précision et rappel. À ce titre, elle est plus pertinente que des indicateurs globaux tels que l'*accuracy* ou la *ROC-AUC*, en particulier lorsque le coût associé aux faux négatifs, c'est-à-dire aux *churners* non détectés, est élevé. La *ROC-AUC* est également reportée à titre complémentaire, afin de situer les résultats par rapport à la littérature, tout en étant interprétée avec prudence dans ce contexte.

Cette évaluation est complétée par des métriques issues de la matrice de confusion, à savoir la précision, le rappel et le *F1-score*. Ce dernier est calculé au seuil optimal déterminé sur le jeu de validation, en maximisant le F1 sur la courbe *precision-recall*, plutôt qu'en fixant un seuil arbitraire. Cette approche permet d'adapter la décision au compromis opérationnel entre couverture des *churners* et précision du ciblage.

Afin d'évaluer la robustesse des performances, des intervalles de confiance à 95 % sont estimés par *bootstrap* (1 000 itérations) sur les métriques calculées sur le jeu de test. Cette démarche permet de quantifier l'incertitude associée aux résultats et d'éviter une interprétation fondée sur une estimation ponctuelle.

Trois analyses complémentaires sont par ailleurs conduites. Une courbe de calibration est d'abord utilisée pour évaluer la fiabilité des probabilités prédites, en comparant celles-ci aux fréquences observées. Cet aspect est particulièrement pertinent dans une perspective opérationnelle, où les probabilités peuvent être utilisées directement pour prioriser des actions de rétention. Une courbe de gain cumulatif est ensuite mobilisée afin de traduire les performances en termes de ciblage client, en mesurant la proportion de *churners* capturés en fonction de la part de la population ciblée.

Enfin, une analyse de robustesse, conduite sur le jeu de test complet, est menée en reformulant le problème en prédiction de la note de satisfaction (*review_score* de 1 à 5). Cette approche permet de vérifier que le classement des méthodes ne dépend pas uniquement du seuil de binarisation retenu. Ainsi, des modèles de régression, à savoir *Ridge* et *XGBoost*, sont utilisés, et le *prompt* de la condition *zero-shot* est adapté en conséquence.

3.6 Conclusion du chapitre

Ce chapitre a présenté le dispositif méthodologique mobilisé pour comparer l'apport de différentes représentations textuelles dans la prédiction du churn. L'analyse porte plus particulièrement sur deux approches supervisées, les embeddings contextuels (M2) et un pipeline RAG (M3), dans le cadre de la prédiction du churn client, en les situant par rapport à deux références, une baseline structurée sans texte (M1) et une approche zero-shot (M4).

En s'appuyant sur un cadre expérimental strictement contrôlé, cette approche permet d'isoler l'effet du type de représentation tout en garantissant la comparabilité des conditions analysées. L'ensemble des choix réalisés, de la construction du dataset à la définition des modèles et des protocoles d'évaluation, vise à assurer la robustesse des résultats et leur interprétation. Ce cadre constitue ainsi la base sur laquelle repose l'analyse des performances présentée dans le chapitre suivant. L'ensemble du code source développé dans le cadre de ce mémoire, du prétraitement des données à la génération des résultats, est disponible en Annexe 9.

Résultats et Analyses

Cette section présente les résultats obtenus pour les différentes conditions expérimentales introduites dans la méthodologie. Elle vise à évaluer l'apport des représentations textuelles dans la prédiction du churn, ainsi qu'à comparer les performances, la stabilité et les caractéristiques des approches étudiées.

L'analyse s'appuie à la fois sur des métriques quantitatives, des tests statistiques ainsi que plusieurs analyses complémentaires afin d'évaluer la robustesse et la portée des résultats obtenus.

4.1 Paramètres empiriques du pipeline RAG

Deux paramètres du pipeline RAG ont fait l'objet d'une validation empirique préalable à l'évaluation principale.

Le premier porte sur le nombre de passages récupérés par FAISS, où l'analyse de sensibilité montre que les corrélations entre scores RAG et note de satisfaction sont stables et élevées pour $k \in \{1,2,3\}$, avec une chute notable à $k = 5$ suggérant une dilution du signal par l'ajout de passages peu pertinents, avec une remontée aux valeurs supérieures. $k = 3$ est retenu comme compromis garantissant une couverture suffisante sans bruit informationnel (voir Annexe 5).

Le second paramètre concerne la taille de l'échantillon d'entraînement de la condition M3. La learning curve mobilisée (voir Annexe 6) met en évidence une stabilisation de la PR-AUC à partir de 1 500 observations, avec une variation marginale inférieure à 0,002 entre 1 500 et 2 000, tant pour la régression logistique que pour XGBoost. L'entraînement est donc conduit sur un échantillon de 1 500 observations, tiré de manière stratifiée sur la variable de churn au sein du jeu d'entraînement.

4.2 Performance globale des conditions expérimentales

Les performances des différentes conditions sont comparées sur le jeu de test indépendant, à partir des meilleures configurations (régression logistique ou XGBoost selon le cas), avec des intervalles de confiance à 95 % obtenus par bootstrap. Le jeu de test est constitué de 5 832 observations, non utilisées pendant la phase d'optimisation. Les performances détaillées de l'ensemble des configurations sont présentées en Annexe 10.

TABLE 2 – Meilleurs modèles par condition : performances sur le jeu de test

Condition	**Modèle**	**PR-AUC**	**IC 95 %**	**F1**	**Recall**	**Précision**
M1	XGB	0,441	$\pm$0,031	0,412	0,413	0,412
M2	LR	0,829	$\pm$0,024	0,801	0,887	0,730
M3	XGB	0,821	$\pm$0,024	0,808	0,889	0,741
M4	Zero-shot	0,816	$\pm$0,022	0,830	0,925	0,754

Le Tableau 2 indique une rupture nette entre la condition M1 et les approches intégrant le texte. Avec une PR-AUC de 0,441 et un F1 de 0,412, M1 reste largement en dessous des autres configurations. À l'inverse, les trois conditions textuelles (M2, M3, M4) se situent toutes au-dessus de 0,80 en PR-AUC. En termes de classement, M2 et M3 occupent les premières positions avec des performances statistiquement équivalentes, suivies de M4 dont le profil se distingue par un recall particulièrement élevé, tandis que M1 reste nettement en retrait. L'écart observé renvoie à une différence structurelle importante, indiquant que les variables transactionnelles seules capturent un signal prédictif limité par rapport à l'information contenue dans les données textuelles. Afin de vérifier que ces écarts ne sont pas attribuables au hasard, des tests de permutation sont conduits entre les conditions.

Apport des données textuelles

TABLE 3 – Tests de permutation

Comparaison	**Δ PR-AUC**	**p-valeur**
M1 XGB $\rightarrow$ M2 LR	$+$0,388	$<$ 0,001
M1 XGB $\rightarrow$ M3 XGB	$+$0,380	$<$ 0,001
M2 LR $\rightarrow$ M3 XGB	$-$0,008	0,749

Les tests de permutation reportés dans le Tableau 3 permettent d'évaluer la significativité des écarts observés à partir d'une distribution nulle construite sur 1 000 permutations des prédictions. Les résultats obtenus confirment la robustesse des différences observées entre conditions.

Le passage de M1 à M2 entraîne un gain de PR-AUC de $+$0,388 ($p < 0{,}001$), de la même manière, le passage de M1 à M3 conduit à un gain comparable de $+$0,380 ($p < 0{,}001$). Dans les deux cas, l'amélioration est à la fois substantielle et statistiquement significative. Ces résultats montrent que l'intégration du texte modifie profondément la capacité prédictive des modèles, au point de rendre insuffisante une approche fondée uniquement sur des données structurées.

M4 n'est pas inclus dans les tests de permutation, ses prédictions n'étant pas issues d'un processus d'optimisation sur les données d'entraînement. La comparaison avec les conditions supervisées repose donc uniquement sur les métriques d'évaluation communes.

Embeddings vs RAG

La comparaison entre M2 et M3 fait apparaître un écart de *PR-AUC* de $-0{,}008$, associé à une p-valeur de 0,749. Cette différence n'est pas statistiquement significative, indiquant des performances équivalentes du point de vue prédictif. Aucune différence significative ne permet donc de départager les deux représentations du texte sur la base de la performance brute.

Cette équivalence ne préjuge toutefois pas de leur interprétabilité respective, qui sera discutée ultérieurement.

Zero-shot

La condition M4 atteint une PR-AUC de 0,816, proche des approches supervisées qui intègrent du texte. Elle se distingue par un F1 (0,830) et un recall (0,925) plus élevés que les autres méthodes. Ce profil traduit une orientation plus forte vers la détection des churners, au détriment d'un équilibre différent entre précision et rappel. Il ne s'agit donc pas d'une amélioration globale des performances, mais d'un profil distinct, dont les implications seront précisées à travers l'analyse de calibration.

4.3 Comparaison à taille d'entraînement fixe entre embeddings et RAG

L'analyse principale compare M2 entraîné sur 27 213 observations à M3 entraîné sur 1 500 observations. Bien qu'elles soient toutes les deux évaluées sur le même jeu de test, ce qui garantit la validité de la comparaison en valeur absolue, l'asymétrie dans les données d'entraînement constitue une limite dans l'interprétation de l'équivalence observée. Afin d'y remédier, M2 est entraîné sur le même sous-échantillon de 1 500 observations que M3, et évaluée sur le même jeu de test.

TABLE 4 – Comparaison de M2 et M3 à taille d'entraînement fixe et jeu de test complet

Condition	**Train**	**PR-AUC**	**IC 95 %**	**ROC-AUC**	**F1**
M2 LR	1 500	0,723	$\pm$0,031	0,927	0,724
M2 XGB	1 500	0,772	$\pm$0,028	0,948	0,753
M3 XGB	1 500	0,821	$\pm$0,024	0,955	0,808

À taille d'entraînement fixe, les résultats du Tableau 4 montrent que M3 XGB (PR-AUC = 0,821) surpasse M2 XGB (PR-AUC = 0,772) ainsi que M2 LR (PR-AUC = 0,723). Le test de permutation confirme que l'écart observé entre M2 XGB et M3 XGB est statistiquement significatif

($\Delta = +0{,}049$, $p < 0{,}001$), ce qui nuance la conclusion d'équivalence formulée dans l'analyse principale. Les embeddings et le RAG ne semblent équivalents qu'à pleine capacité des embeddings. À taille d'entraînement identique, le pipeline RAG présente un avantage prédictif net, ce qui montre que sa représentation textuelle structurée est plus efficiente par observation.

Ces deux lectures sont complémentaires. L'analyse à pleine capacité reflète les performances des embeddings dans des conditions favorables en volume de données, tandis que la comparaison à taille fixe permet d'isoler l'effet propre de la représentation textuelle. Ensemble, elles indiquent que le RAG est plus efficient à faible volume, alors que les embeddings peuvent combler l'écart lorsqu'ils disposent de davantage de données.

4.4 Diagnostic d'overfitting

L'analyse de la stabilité des modèles repose sur l'écart entre les performances sur le jeu d'entraînement et le jeu de validation. L'objectif est de vérifier si les performances observées sont robustes ou si elles traduisent un phénomène d'overfitting.

Les résultats montrent que les modèles ne présentent pas tous le même comportement, les écarts variant selon les configurations expérimentales. La configuration la plus stable est celle de M2 avec la régression logistique. L'écart y est faible, de l'ordre de 0,049, ce qui indique une bonne capacité de généralisation entre le jeu d'entraînement et le jeu de validation, sans surapprentissage significatif. M2 XGBoost présente quant à lui un écart légèrement plus élevé (0,097), tout en restant dans une plage acceptable.

Pour M1, l'écart atteint 0,148 pour LR et 0,175 pour XGB. Ces valeurs ne traduisent pas un surapprentissage classique corrigeable par régularisation, mais reflètent la limite intrinsèque des features transactionnelles face au drift temporel du taux de churn entre le jeu d'entraînement et de validation.

La situation est différente pour M3, où le gap est également présent, mais plus faible, compris entre 0,073 (LR) et 0,085 (XGBoost). Celui-ci s'explique principalement par la taille réduite de l'échantillon d'entraînement, limité à 1 500 observations. L'écart observé ne reflète pas un défaut de paramétrage mais une contrainte d'échantillonnage inhérente au coût du pipeline RAG.

Dans les deux cas, ces écarts sont interprétés comme structurels et documentés comme limites du dispositif expérimental.

4.5 Feature importance

Cette section examine les variables qui orientent les prédictions afin d'identifier la nature du signal de churn dans le dataset Olist. L'analyse se concentre sur les modèles M1 XGB et M3 XGB, pour lesquels les mesures d'importance des variables sont disponibles. L'objectif est de déterminer si, malgré des modes de représentation textuels distincts, les deux approches convergent vers les mêmes déterminants du churn.

Dans le cas de M1 XGB (Figure 2), la variable `is_late` apparaît comme largement dominante, avec une importance de 0,648. Elle est suivie par le nombre d'articles de la commande (`n_items`, 0,133) et du délai de retard (`delivery_delay_days`, 0,078). Les deux variables liées à la livraison cumulent à elles seules 72,6 % de l'importance totale, ce qui indique que dans une approche fondée uniquement sur les données transactionnelles, le churn est principalement associé à des défaillances logistiques.

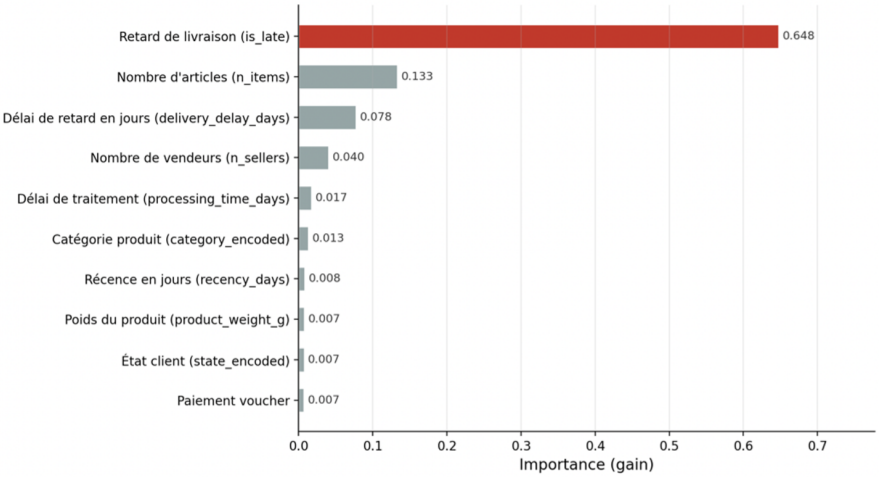


FIGURE 2 – Feature importance – M1 XGB

Dans le cas de M3 XGB (Figure 3), le score RAG associé à la livraison constitue la variable la plus importante (0,253). La variable `is_late`, qui occupait la première position dans M1, recule ici au quatrième rang. Elle est précédée par deux autres variables issues du pipeline RAG, correspondant à la qualité produit (0,161) et au service client (0,105). Ce réagencement des importances ne remet toutefois pas en cause la prépondérance de la livraison, l'ensemble des variables qui lui sont associées concentrant 50,5 % de l'importance totale du modèle.

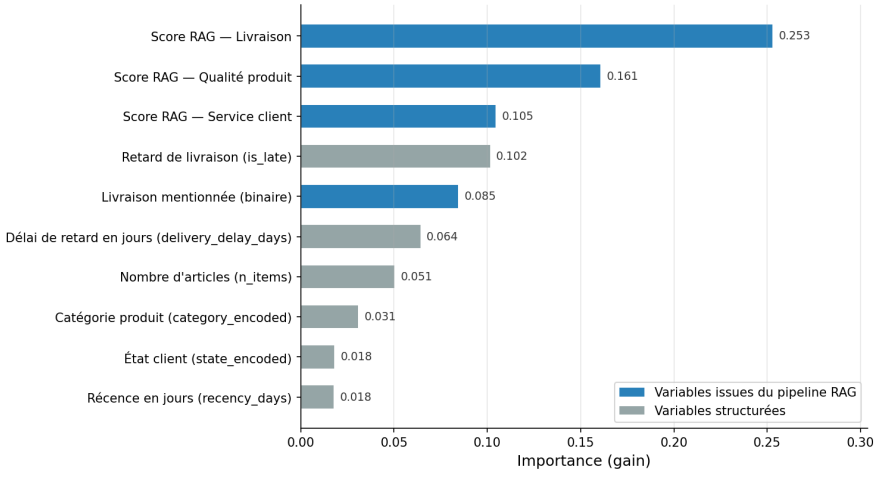


FIGURE 3 – Feature importance – M3 XGB

Malgré la différence de nature entre données structurées et scores sémantiques issus du RAG, les deux approches convergent vers un même déterminant principal, la livraison. Cette convergence

dépasse la seule comparaison des performances et renforce la cohérence interprétative des résultats, en montrant que des méthodes distinctes aboutissent à une même lecture du phénomène étudié.

Les résultats détaillés des importances de variables pour M1 XGB et M3 XGB sont présentés en Annexe 11.

4.6 Analyses complémentaires

4.6.1 Calibration

L'évaluation des modèles ne se limite pas à leur performance discriminante. Deux modèles présentant une PR-AUC similaire peuvent produire des probabilités de nature très différente. La calibration permet précisément d'évaluer dans quelle mesure les probabilités prédites correspondent aux fréquences observées. À titre d'exemple, un modèle bien calibré attribuant une probabilité de 0,80 à un client implique que, parmi l'ensemble des clients recevant ce score, environ 80 % sont effectivement des churners. Cette propriété conditionne directement leur utilisation dans un contexte opérationnel, notamment lorsqu'elles servent à prioriser des actions de rétention.

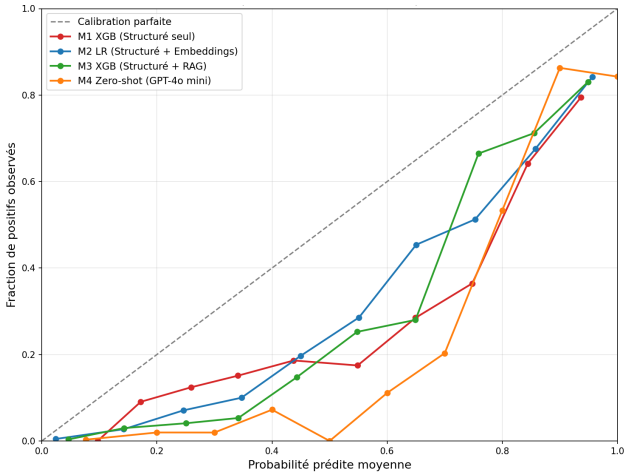


FIGURE 4 – Courbes de calibration des quatre conditions expérimentales

Les résultats montrent des écarts marqués entre les différentes conditions (Figure 4). M2 LR est la condition la mieux calibrée, ce qui est conforme aux propriétés de la régression logistique, dont la fonction sigmoïde favorise une estimation probabiliste cohérente. M1 XGB présente une calibration acceptable dans les faibles probabilités, avec une tendance à sous-estimer dans la zone intermédiaire, entre 0,4 et 0,8. M3 XGB présente un comportement irrégulier dans les faibles probabilités, sa courbe s'écartant de manière non monotone de la diagonale idéale. Ce phénomène pourrait être attribué à la concentration des scores RAG autour de 0,5 pour les dimensions rarement mentionnées, ce qui conduit à des intervalles de faibles probabilités contenant peu d'observations.

Le modèle M4 zero-shot se distingue, quant à lui, par une calibration nettement dégradée. Les probabilités prédites sont fortement concentrées aux extrêmes, ce qui traduit une tendance à pro-

duire des prédictions très tranchées. Ce comportement est cohérent avec le recall élevé observé précédemment, mais s'accompagne d'une faible correspondance entre probabilités prédites et fréquences réelles.

Il convient toutefois de souligner que cette limite porte uniquement sur l'interprétation directe des probabilités brutes, et non sur la capacité discriminante des modèles, qui reste évaluée indépendamment à travers la PR-AUC.

4.6.2 Courbe de gain cumulatif

L'analyse des performances est ici transposée dans une perspective opérationnelle, en évaluant la capacité des modèles à prioriser les clients à risque. L'objectif est de mesurer la proportion de churners effectivement capturés en fonction du pourcentage de clients ciblés, dans une logique de campagnes de rétention.

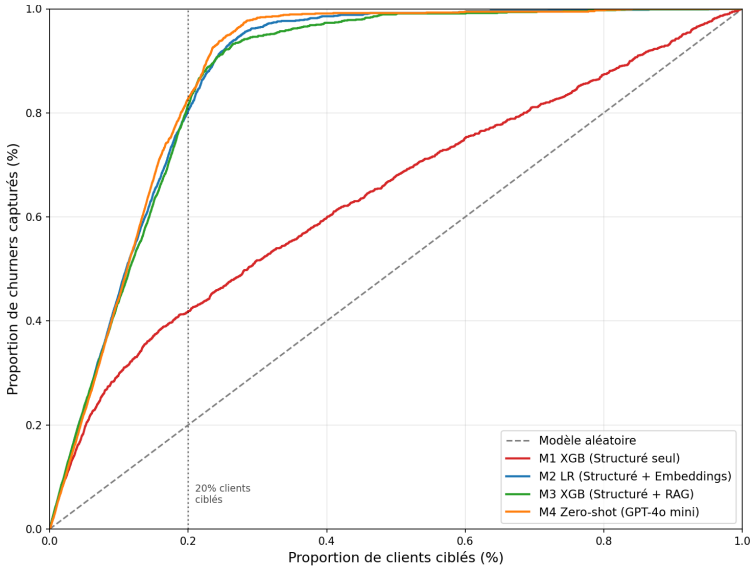


FIGURE 5 – Courbes de gain cumulatif

Les résultats mettent en évidence un écart net entre la condition fondée uniquement sur les données structurées et celles intégrant le texte (Figure 5). Cet écart apparaît dès les premiers déciles, traduisant une capacité nettement supérieure des modèles textuels à identifier les churners prioritaires. Par ailleurs, les courbes associées aux conditions M2, M3 et M4 se superposent largement, indiquant des performances similaires en termes de ciblage.

TABLE 5 – Proportion de churners capturés selon le seuil de ciblage

Méthode	**10 %**	**20 %**	**30 %**	**50 %**
M1 XGB	29,7 %	41,9 %	51,7 %	67,9 %
M2 LR	45,2 %	80,4 %	96,3 %	99,2 %
M3 XGB	43,9 %	81,5 %	94,6 %	98,9 %
M4 ZS	44,4 %	82,8 %	98,3 %	99,2 %

Les résultats détaillés confirment cette tendance (Tableau 5). En ciblant 20 % des clients les plus à risque, les modèles M2, M3 et M4 permettent de capturer environ 80 % des churners, contre seulement 41,9 % pour M1. Lorsque le seuil de ciblage est porté à 30 %, la proportion de churners capturés dépasse 94 % pour les données textuelles, atteignant jusqu'à 98,3 % pour M4. À l'inverse, M1 reste significativement en retrait à tous les niveaux de ciblage.

4.7 Analyse de robustesse

Depuis le début de l'analyse, la définition du churn repose sur un seuil de binarisation (≤ 2), ce qui introduit une dépendance potentielle aux choix méthodologiques. L'objectif de cette section est d'évaluer si les résultats obtenus restent valides en l'absence de cette binarisation.

À cette fin, une seconde configuration est introduite, dans laquelle la variable cible correspond directement à la note de satisfaction brute (*review_score* de 1 à 5). Cette approche transforme ainsi le problème en une tâche de prédiction ordinale. En conséquence, les métriques d'évaluation sont adaptées. En effet, la corrélation de Spearman est retenue comme métrique principale, dans la mesure où elle évalue la capacité du modèle à préserver l'ordre relatif des observations, indépendamment des écarts absolus.

TABLE 6 – Analyse de robustesse

Condition	**Corrélation de Spearman**	**Classement**
M1	0,245	4^{e}
M2	0,709	2^{e}
M3	0,676	3^{e}
M4	0,809	1er

Les résultats du Tableau 6 mettent en évidence une stabilité globale du classement des méthodes entre les deux configurations. M1 conserve la dernière position, confirmant que l'avantage des

approches textuelles est indépendant du seuil de binarisation. Le classement interne entre conditions textuelles évolue toutefois légèrement. M4, quatrième dans l'analyse principale, progresse vers la première position dans cette analyse, ce qui suggère que le modèle zero-shot est particulièrement adapté à la prédiction ordinale continue.

Des tests de Wilcoxon ont également été réalisés afin d'évaluer si les écarts observés entre méthodes restent statistiquement significatifs à l'échelle individuelle. Toutes les comparaisons produisent des p-valeurs inférieures à 0,001, indiquant que les différences d'erreurs absolues entre conditions ne relèvent donc pas du hasard. Il convient néanmoins de noter que la significativité observée entre M2 et M3 dans ce test de robustesse contraste avec l'absence de significativité dans l'analyse principale, où M2 est entraîné sur le jeu d'entraînement complet ($p = 0{,}749$ au test de permutation). Ce résultat suggère que les deux approches se distinguent davantage sur la prédiction ordinale de la note brute que sur la classification binaire du churn, ce qui reste cohérent avec le caractère continu de la variable cible.

Par ailleurs, les performances de M4 sont particulièrement élevées (Spearman = 0,809), ce qui traduit une forte capacité à ordonner correctement les observations. Cette performance doit toutefois être interprétée avec prudence. M4 produit nativement une probabilité continue entre 0 et 1, tandis que M2 et M3 sont optimisés pour une cible binaire via PR-AUC. La tâche de prédiction ordinale avantage donc structurellement M4, dont les sorties sont naturellement adaptées à un classement continu. Sa première position dans ce test de robustesse doit ainsi être interprétée comme un avantage lié à la nature de la tâche, et non comme une supériorité générale sur les approches supervisées.

En définitive, cette analyse confirme que la hiérarchie des méthodes est robuste au choix du seuil de binarisation. Les métriques détaillées (MAE, RMSE, Spearman et tests de Wilcoxon) sont reportées en Annexe 12.

4.8 Conclusion du chapitre

Les résultats montrent d'abord que l'intégration des données textuelles améliore fortement la prédiction du churn par rapport aux données structurées seules. La condition M1 reste nettement en dessous des approches textuelles, confirmant que le signal de churn dans Olist est principalement porté par les avis clients. Les embeddings contextuels et le pipeline RAG atteignent des performances globalement équivalentes à pleine capacité des embeddings. Toutefois, à taille d'entraînement fixe, M3 surpasse significativement M2, suggérant une meilleure efficience du RAG par observation. Les analyses de feature importance convergent vers la livraison comme principal déterminant du churn, quelle que soit la méthode utilisée. Enfin, le modèle zero-shot obtient des performances compétitives sans entraînement supervisé, tandis que l'analyse de robustesse confirme la stabilité globale des résultats.

Ces observations permettent de comparer les performances des différentes approches, avant d'en discuter les implications opérationnelles et managériales dans la section suivante.

Discussion

Ce chapitre présente l'interprétation des résultats obtenus et prolonge l'analyse empirique en passant de l'évaluation des performances à leur mise en perspective au regard de la question de recherche. L'enjeu est de comprendre dans quelle mesure les différentes représentations textuelles mobilisées contribuent à la prédiction du churn client, mais également d'identifier les mécanismes sous-jacents aux performances observées, notamment à travers l'analyse des signaux les plus déterminants. Pour structurer cette discussion, l'analyse s'articule autour de l'interprétation globale des résultats, de la lecture des déterminants du churn, puis des implications méthodologiques et managériales ainsi que des limites et pistes de recherche.

5.1 Interprétation des résultats au regard de la question de recherche

Les résultats obtenus montrent tout d'abord que l'intégration des reviews textuelles améliore fortement la prédiction du churn par rapport à une approche fondée uniquement sur des données structurées. Les tests statistiques confirment que ce gain est robuste et non attribuable au hasard. Les variables transactionnelles seules apparaissent donc insuffisantes pour capturer l'ensemble du signal de churn dans le contexte d'Olist. À l'inverse, les avis clients apportent une information comportementale complémentaire, liée à l'expérience vécue par le client, difficilement observable à travers les seules données tabulaires. Ce résultat confirme les travaux de la littérature soulignant la forte valeur prédictive des données textuelles dans les tâches de churn et d'analyse de satisfaction client (De Caigny et al., 2020).

Dès lors, la question n'est plus réellement de savoir s'il est pertinent d'intégrer du texte dans les modèles de prédiction, mais plutôt de déterminer comment ce texte doit être représenté et exploité.

À ce niveau, les résultats mettent en évidence une relation nuancée entre les embeddings et le RAG. Dans l'analyse principale, les deux approches présentent des performances statistiquement équivalentes, mais lorsque la comparaison est effectuée à taille d'échantillon d'entraînement identique, le modèle basé sur le RAG surpasse celui reposant sur les embeddings.

Ce résultat s'interprète comme un avantage d'efficience pour M3, qui parvient à extraire davantage d'information prédictive par observation que les embeddings contextuels en M2. Cette propriété s'explique par la nature même du RAG, qui ancre l'extraction dans des dimensions théoriques explicites. Les embeddings, en revanche, ne retrouvent des performances comparables que lorsqu'ils disposent d'un volume de données d'entraînement suffisant pour compenser leur absence de structuration préalable.

Cette proximité en performance ne doit toutefois pas être interprétée comme une équivalence

méthodologique. Les embeddings utilisés dans M2 reposent sur une représentation vectorielle dense, difficilement interprétable et fonctionnant comme boîte noire. À l'inverse, le pipeline RAG produit des dimensions explicites, alignées avec des déterminants théoriques du churn identifiés dans la littérature. Sa valeur ne réside donc pas uniquement dans la performance prédictive, mais également dans la production de variables interprétables et directement exploitables dans une logique analytique et métier.

Un résultat particulièrement notable concerne l'approche zero-shot. Malgré l'absence totale d'apprentissage supervisé, M4 atteint un niveau de performance comparable aux approches supervisées, avec un profil distinct caractérisé par un recall particulièrement élevé et de meilleures performances en analyse ordinale, en cohérence avec la nature continue de la tâche. Ce résultat illustre la capacité actuelle des modèles de langage à extraire directement un signal comportemental complexe à partir du texte brut.

En définitive, les résultats appellent une interprétation plus fine du signal identifié, notamment quant à la nature du déterminant principal du churn dans le contexte d'Olist.

5.2 Le signal de churn dans Olist

L'analyse des feature importance des modèles XGBoost permet d'aller au-delà de la simple comparaison des performances prédictives et apporte une lecture analytique et métier des signaux qui déterminent concrètement le churn dans Olist.

Comme l'indiquent les résultats, les méthodes M1 et M3 convergent vers un même déterminant principal. Dans M1, la variable `is_late` domine très largement le classement des features les plus importantes, suivie par plusieurs variables liées à la logistique et aux délais de livraison, représentant à elles seules 72,6 % de l'importance totale du modèle.

Dans le cadre de M3, le score RAG, l'indicateur binaire de mention et le délai de retard cumulent 50,5 % de l'importance, contre 16,1 % pour la qualité produit.

Ainsi, malgré des représentations du texte fondamentalement différentes, les deux approches convergent vers le même signal dominant. Ce constat renforce la robustesse analytique des résultats et suggère qu'il ne s'agit pas d'un simple artefact méthodologique lié à une représentation particulière des données.

Les résultats indiquent donc que le churn dans Olist semble principalement lié à l'expérience logistique, ce qui reste cohérent avec la littérature sur le e-commerce dans laquelle plusieurs travaux ont déjà identifié la livraison comme un déterminant majeur de satisfaction et de fidélisation client (Archak et al., 2011 ; Tirunillai & Tellis, 2014). Les résultats de ce mémoire s'inscrivent ainsi dans le prolongement de ces observations sur données réelles.

À l'inverse, les dimensions liées au prix et au service client apparaissent beaucoup moins discriminantes. Ce résultat est lui-même informatif dans la mesure où les commentaires clients ne reflètent pas avec la même intensité l'ensemble des déterminants du churn, certaines dimensions étant plus rarement exprimées dans les avis textuels.

5.3 Le zero-shot comme référence non supervisée

M4 occupe une position particulière dans cette étude, dans la mesure où cette approche fonctionne sans entraînement supervisé ni feature engineering. Son profil distinct justifie ainsi une discussion spécifique.

Dans l'analyse principale, l'approche zero-shot atteint des performances comparables à celles des méthodes supervisées M2 et M3. Plus encore, elle obtient les meilleurs résultats dans l'analyse de robustesse fondée sur la prédiction ordinale des notes, ce qui s'explique par son avantage structurel. Ce résultat est d'autant plus marquant qu'il positionne M4 au niveau des approches supervisées entraînées sur des milliers d'observations étiquetées. Il suggère que les modèles de langage modernes sont capables d'extraire directement, à partir de texte brut, des signaux comportementaux complexes liés au churn. Leur compréhension sémantique semble ainsi naturellement adaptée à l'interprétation de signaux continus et nuancés. Plus largement, ce résultat illustre le potentiel des modèles généralistes à rivaliser avec des pipelines supervisés plus lourds.

Ces résultats doivent toutefois être interprétés avec prudence, les analyses de calibration ayant montré que les probabilités produites par M4 restent peu fiables sans recalibration préalable. Dans une perspective métier, M4 semble donc davantage adapté à des stratégies de rétention à large spectre visant à maximiser le rappel et à identifier un maximum de churners potentiels, plutôt qu'à des contextes nécessitant un scoring précis et calibré.

5.4 Implications managériales

Les résultats obtenus permettent de formuler plusieurs recommandations concrètes pour les plateformes d'e-commerce. L'enjeu n'est pas uniquement prédictif, mais également opérationnel puisque ces modèles peuvent orienter des décisions marketing, logistiques et même stratégiques.

Recommandation 1 : intégrer les avis textuels dans les modèles du churn

Les analyses ont mis en évidence que l'écart entre M1 et les conditions textuelles est massif et statistiquement significatif. Une approche fondée uniquement sur les données transactionnelles apparaît donc insuffisante dans un contexte où des avis clients sont disponibles.

Les reviews textuelles capturent des signaux comportementaux absents des données tabulaires classiques. Elles permettent notamment d'identifier des formes d'insatisfactions avant qu'elles ne se traduisent concrètement dans les comportements d'achat ou de départ. Dans cette perspective, les avis clients ne doivent plus être considérés comme une source d'information secondaire, mais comme une composante centrale des systèmes de prédiction du churn.

Recommandation 2 : choisir entre embeddings et RAG selon l'objectif métier

Au regard des performances observées entre M2 et M3, le choix entre embeddings et RAG dépend principalement du contexte d'usage et des besoins métiers.

Lorsque l'objectif principal est la performance brute avec des ressources techniques limitées, les embeddings contextuels constituent l'option la plus pertinente. Le pipeline est plus simple à mettre en œuvre, moins coûteux et plus rapide à déployer. Aucun appel API n'est nécessaire, les représentations pouvant être générées localement à partir de modèles préentraînés. Cette approche est donc particulièrement adaptée à des systèmes de scoring automatisés à grande échelle.

À l'inverse, lorsque l'objectif inclut une dimension d'interprétabilité ou de compréhension métier, le pipeline RAG apporte une valeur ajoutée importante. Les variables produites sont directement interprétables et permettent d'identifier explicitement les dimensions associées au churn. Cette structure facilite ensuite la communication avec les équipes marketing ou opérationnelles, qui peuvent relier les prédictions à des problématiques concrètes et actionnables.

Par ailleurs, dans des contextes où le volume de données étiquetées est limité, le pipeline RAG constitue l'option à privilégier au vu de son avantage prédictif net à taille d'entraînement réduite.

Ainsi, l'intérêt principal du RAG réside également dans l'exploitabilité analytique des résultats. Le choix entre les deux approches correspond donc davantage à un arbitrage entre simplicité technique et richesse interprétative, ainsi qu'à un compromis potentiel entre performance et contrainte de données selon les contextes d'application.

Recommandation 3 : prioriser la livraison comme levier de rétention

Les analyses de feature importance menées sur M1 et M3 convergent vers un même constat qui identifie la livraison comme le principal déterminant du churn dans Olist. Cette convergence apparaît dans plusieurs méthodes indépendantes reposant sur des représentations très différentes des données, ce qui renforce la robustesse du résultat.

Dans une perspective opérationnelle, plusieurs actions peuvent être envisagées : réduire les délais de livraison, améliorer la fiabilité logistique, renforcer la transparence sur les délais annoncés ou encore accélérer la gestion des incidents de livraison, l'objectif étant de limiter le risque de départ associé à ces dysfonctionnements.

Recommandation 4 : prévoir une étape de recalibration avant tout déploiement opérationnel

Les probabilités produites par les différentes approches présentent des écarts avec les fréquences réellement observées, ce qui limite leur utilisation directe dans un contexte de scoring client. Cette problématique est particulièrement marquée pour M4, dont la calibration est la plus dégradée, tandis que M3 présente un comportement irrégulier sur les faibles probabilités et que M2 apparaît globalement mieux calibré, bien que perfectible.

Dans un cadre opérationnel, une étape de recalibration est donc nécessaire avant tout déploiement. Des méthodes comme le *Platt scaling* ou l'*isotonic regression* permettent précisément d'ajus-

ter les probabilités produites afin qu'elles reflètent mieux le risque observé dans les données. L'objectif est donc de rendre les scores de l'ensemble des conditions exploitables pour la prise de décision opérationnelle.

Recommandation 5 : encadrer le recours au zero-shot

Le zero-shot présente un intérêt réel dans certains contextes. Son rappel élevé le rend pertinent pour des stratégies de rétention à large spectre, où l'objectif est de détecter un maximum de clients à risque, quitte à accepter davantage de faux positifs.

Il constitue également une option intéressante dans des contextes exploratoires ou lorsque peu de données annotées sont disponibles.

Son recours suppose toutefois de prendre en compte une contrainte structurelle. L'utilisation d'un LLM externe introduit une dépendance à un modèle propriétaire dont les sorties peuvent varier selon les versions ou les mises à jour, ce qui limite la stabilité du pipeline dans le temps. Son utilisation implique donc d'adapter son recours aux contraintes du cas d'usage et au niveau d'exigence attendu en production.

5.5 Considérations éthiques

Au-delà des performances techniques, le déploiement de tels systèmes soulève plusieurs questions éthiques qui méritent d'être explicitées.

En premier lieu, l'utilisation de GPT-4o mini dans les conditions M3 et M4 introduit une dépendance à un modèle propriétaire dont les mécanismes internes ne sont pas accessibles. Cette opacité rend difficile l'identification de biais potentiels, notamment culturels ou linguistiques, dans l'interprétation des avis en portugais. Le recours à des modèles open-source fine-tunés sur des corpus locaux constituerait une alternative plus transparente et auditable.

En second lieu, étiqueter un client comme étant à risque de churn à partir de ses avis constitue une forme de profilage comportemental. Si ce profilage reste interne à l'entreprise, il soulève la question de la transparence vis-à-vis des clients ciblés par des actions de rétention. Des pratiques telles que l'information préalable des clients ou la mise en place d'un droit d'opposition permettraient d'encadrer ces usages.

Enfin, les avis textuels mobilisés sont des données personnelles au sens du RGPD, et leur intégration dans un pipeline prédictif suppose une base légale explicite ainsi qu'une politique de conservation adaptée. Ces considérations rappellent que la performance prédictive ne suffit pas à légitimer un déploiement opérationnel et qu'un cadre éthique défini constitue un prérequis à toute mise en production.

5.6 Limites

Malgré les contributions apportées par ce mémoire, plusieurs limites doivent être prises en compte dans l'interprétation des résultats. Ces limites concernent à la fois les données mobilisées, le design expérimental adopté ainsi que les modèles utilisés.

Limites liées aux données

La principale limite de cette étude concerne la définition même du churn. Comme détaillé dans la partie méthodologique, l'analyse repose sur un proxy d'insatisfaction basé sur les scores de reviews en l'absence de données longitudinales permettant d'observer un churn comportemental réel. Les résultats obtenus doivent donc être interprétés comme une prédiction d'insatisfaction fortement associée au churn, et non comme une mesure stricte du départ client.

Par ailleurs, un biais de sélection des clients expressifs est également présent. Dans les conditions textuelles, seules les observations contenant des reviews exploitables peuvent être utilisées. Pourtant, les clients qui rédigent des commentaires ne sont pas nécessairement représentatifs de l'ensemble de la clientèle. Certains profils ont davantage tendance à exprimer leur satisfaction ou leur insatisfaction publiquement, ce qui peut introduire un biais dans les signaux appris par les modèles.

Enfin, l'étude repose sur un unique dataset issu du contexte du e-commerce brésilien. Les résultats obtenus ne peuvent donc pas être généralisés automatiquement à d'autres secteurs, environnements ou culturels.

Limites liées au design expérimental

Le pipeline RAG repose sur un corpus limité à 41 passages, construits manuellement et spécialisés autour de dimensions précises du churn. Bien que cette approche permette une meilleure interprétabilité, elle peut également limiter la richesse des signaux sémantiques extraits. Par ailleurs, le choix des quatre dimensions retenues, (livraison, qualité produit, service client et prix), bien qu'ancré dans la littérature, ne couvre pas l'ensemble des déterminants potentiels du churn en e-commerce. D'autres dimensions plus discriminantes pourraient être mobilisées. Parmi celles-ci, l'expérience d'achat sur le site, le processus de retour ou encore la communication post-achat constituent des pistes d'enrichissement du pipeline non explorées dans ce travail.

En outre, aucune baseline textuelle lexicale (type TF-IDF) n'a été intégrée au dispositif expérimental. Sans cette condition intermédiaire, il n'est pas possible d'isoler ce qui relève du caractère sémantique et dense des embeddings de la simple présence d'information lexicale dans le modèle. Une partie du gain observé entre M1 et M2 pourrait provenir d'un signal lexical que TF-IDF aurait également capturé.

De plus, le modèle M3 n'a été entraîné que sur 1 500 observations, principalement en raison des contraintes de coût et de temps associées aux appels API. Cette taille réduite du jeu d'entraînement

peut limiter la capacité de généralisation du modèle.

Ces contraintes ont notamment contribué à l'apparition d'un overfitting structurel dans certaines conditions expérimentales, particulièrement pour M1 et M3. Cet overfitting ne résulte pas d'un mauvais réglage des modèles, mais plutôt des limites inhérentes aux données et au protocole expérimental. Dans ce contexte, la régularisation seule ne suffisait pas à corriger complètement les écarts observés entre entraînement et validation.

Limites liées aux modèles

Les conditions M3 et M4 reposent sur l'utilisation de GPT-4o mini, ce qui introduit une dépendance à un modèle propriétaire externe dont l'évolution dans le temps n'est pas maîtrisable.

Enfin, bien que M2 atteigne des performances élevées, les embeddings utilisés restent largement opaques. Les vecteurs denses générés sont difficiles à relier à des dimensions métier explicites, ce qui limite la compréhension des mécanismes décisionnels du modèle et réduit son interprétabilité opérationnelle.

Ces limites n'invalident pas les résultats obtenus, mais elles imposent une interprétation prudente des conclusions formulées.

5.7 Pistes de recherches futures

Plusieurs pistes de recherche peuvent être envisagées afin de prolonger ce travail et d'en renforcer la portée empirique et méthodologique.

Premièrement, il serait pertinent d'évaluer le pipeline RAG sur un dataset contenant un churn comportemental réel, basé sur des données longitudinales et un historique d'achat complet. Cela permettrait de vérifier si les résultats observés ici se maintiennent lorsque la variable cible correspond directement au non-retour effectif du client, et non à un proxy d'insatisfaction construit à partir des scores de review.

Deuxièmement, l'exploration de modèles locaux fine-tunés sur le portugais représenterait une alternative intéressante aux modèles propriétaires utilisés dans cette étude. Une telle approche permettrait de réduire la dépendance à GPT-4o mini, de limiter les coûts liés aux appels API et d'améliorer la reproductibilité des expériences dans le temps. Elle pourrait également offrir un meilleur contrôle sur les comportements du modèle et sur les données traitées.

Troisièmement, deux pistes complémentaires permettraient d'étendre la portée des résultats. La première consisterait à enrichir la base de connaissances du pipeline RAG avec des dimensions supplémentaires afin de couvrir un spectre plus large des déterminants du churn. La seconde viserait à répliquer l'analyse sur des datasets issus d'autres secteurs ou contextes culturels, afin de déterminer si la prédominance du signal livraison constituent des résultats spécifiques à Olist ou des tendances plus générales.

Par ailleurs, l'intégration d'une baseline lexicale de type TF-IDF permettrait d'isoler plus finement la contribution sémantique des embeddings contextuels par rapport à une simple représentation fondée sur la fréquence des termes. Cette piste constituerait dès lors un prolongement direct du dispositif expérimental.

Enfin, une dernière piste concerne la recalibration des probabilités prédites. Cette étude montre que M4 présente la calibration la plus dégradée, mais M3 et M2 restent également perfectibles. Une évaluation comparative des méthodes de recalibration post-hoc sur les trois modèles constituerait un prolongement direct de ce travail, avec pour objectif d'améliorer l'exploitabilité opérationnelle des scores produits dans un contexte de déploiement réel.

5.8 Conclusion du chapitre

Ces analyses montrent qu'un modèle performant sur les métriques de classification n'est pas nécessairement le plus adapté à un déploiement opérationnel. Le choix d'une approche dépend avant tout du contexte métier, des ressources disponibles et du niveau d'interprétabilité recherché, autant d'éléments qui dépassent la seule comparaison des performances.

Conclusion Générale

Ce mémoire s'est inscrit dans une volonté de répondre à une question méthodologique encore peu explorée dans la littérature sur la prédiction du churn. Dans cette perspective, ce travail propose notamment une évaluation originale du Retrieval-Augmented Generation (RAG) utilisé non plus comme un outil génératif, mais comme un mécanisme d'extraction sémantique structurée destiné à produire des variables exploitables par un modèle prédictif supervisé.

En comparant deux références, une baseline structurée sans texte et un LLM zero-shot, à deux approches supervisées au cœur de l'analyse, les embeddings contextuels et le pipeline RAG, dans un cadre où seul le type de représentation textuelle variait, ce travail apporte des éléments de réponse concrets à cette question.

Les résultats montrent que le pipeline RAG améliore significativement la prédiction du churn par rapport aux données structurées seules, et présente un avantage prédictif net sur les embeddings contextuels lorsque le volume de données d'entraînement est limité. À volume équivalent, les deux approches textuelles atteignent des performances statistiquement équivalentes, mais le RAG conserve un avantage décisif en termes d'interprétabilité et d'efficience par observation. Le profil du modèle zero-shot mérite également d'être souligné. Sans aucun apprentissage supervisé, M4 atteint des performances comparables aux approches supervisées, illustrant les capacités actuelles des grands modèles de langage à extraire directement un signal comportemental complexe à partir du texte brut.

Au-delà des métriques, l'analyse des *feature importance* apporte une lecture analytique cohérente, la livraison émergeant, quelle que soit la méthode mobilisée, comme le principal déterminant du churn dans Olist.

Ces résultats s'appuient sur une contribution méthodologique directement transposable. Un cadre expérimental contrôlé pour comparer des approches hétérogènes de traitement du texte, une validation empirique des variables issues du RAG avant leur intégration prédictive, et une démonstration que la richesse interprétative d'une architecture peut constituer une valeur ajoutée indépendante de sa performance brute.

Plusieurs limites encadrent toutefois l'interprétation des résultats. La variable cible repose sur un proxy d'insatisfaction et non sur un churn comportemental observé, le dataset est spécifique au contexte brésilien, et le pipeline RAG mobilise un modèle propriétaire dont l'évolution dans le temps n'est pas maîtrisable. Ces limites ouvrent autant de pistes de recherche futures, notamment l'évaluation du pipeline sur des données longitudinales réelles ou son extension à d'autres contextes culturels et sectoriels.

Bibliographie

AL-AYED, S. (2022). The impact of e-commerce drivers on e-customer loyalty : Evidence from KSA. *International Journal of Data and Network Science*, *6*(1), 73-80. https://doi.org/10.5267/j.ijdns.2021.10.002

ANDERSON, E. W., FORNELL, C., & MAZVANCHERYL, S. K. (2004). Customer Satisfaction and Shareholder Value. *Journal of Marketing*, *68*(4), 172-185. https://doi.org/10.1509/jmkg.68.4.172.42723

ARCHAK, N., GHOSE, A., & IPEIROTIS, P. G. (2011). Deriving the Pricing Power of Product Features by Mining Consumer Reviews. *Management Science*, *57*(8), 1485-1509. https://doi.org/10.1287/mnsc.1110.1370

ASAI, A., WU, Z., WANG, Y., SIL, A., & HAJISHIRZI, H. (2023). Self-RAG : Learning to Retrieve, Generate, and Critique through Self-Reflection [arXiv :2310.11511]. https://doi.org/10.48550/arXiv.2310.11511

BALDUCCI, B., & MARINOVA, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science*, *46*(4), 557-590. https://doi.org/10.1007/s11747-018-0581-x

BLEI, D. M., NG, A. Y., & JORDAN, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

BUCKINX, W., & VAN DEN POEL, D. (2005). Customer base analysis : Partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research*, *164*(1), 252-268. https://doi.org/10.1016/j.ejor.2003.12.010

CHAJIA, M., & NFAOUI, E. H. (2024). Customer Churn Prediction Approach Based on LLM Embeddings and Logistic Regression. *Future Internet*, *16*(12), 453. https://doi.org/10.3390/fi16120453

COUSSEMENT, K., & VAN DEN POEL, D. (2008). Integrating the voice of customers through call center emails into a decision support system for churn prediction. *Information & Management*, *45*(3), 164-174. https://doi.org/10.1016/j.im.2008.01.005

DE CAIGNY, A., COUSSEMENT, K., DE BOCK, K. W., & LESSMANN, S. (2020). Incorporating textual information in customer churn prediction models based on a convolutional neural network. *International Journal of Forecasting*, *36*(4), 1563-1578. https://doi.org/10.1016/j.ijforecast.2019.03.029

DEVLIN, J., CHANG, M.-W., LEE, K., & TOUTANOVA, K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1*, 4171-4186. https://doi.org/10.18653/v1/N19-1423

FADER, P. S., & HARDIE, B. G. (2009). Probability models for customer-base analysis. *Journal of Interactive Marketing*, *23*(1), 61-69.

FIGALIST, I., ELSNER, C., BOSCH, J., & OLSSON, H. H. (2019). Customer churn prediction in B2B contexts. *International Conference on Software Business*, 378-386.

GANDOMI, A., & HAIDER, M. (2015). Beyond the hype : Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

GAO, Y., XIONG, Y., GAO, X., JIA, K., PAN, J., BI, Y., & WANG, H. (2024). Retrieval-Augmented Generation for Large Language Models : A Survey [arXiv :2312.10997]. https://doi.org/10.48550/arXiv.2312.10997

GHOSH, M. (2024). Meta-analytic review of online purchase intention : Conceptualising the study variables. *Cogent Business & Management*, *11*(1), 2296686.

GUPTA, S., & ZEITHAML, V. (2006). Customer Metrics and Their Impact on Financial Performance.

IMANI, M., JOUDAKI, M., BEIKMOHAMMADI, A., & ARABNIA, H. R. (2025). Customer churn prediction : A systematic review of recent advances, trends, and challenges in machine learning and deep learning. *Machine Learning and Knowledge Extraction*, *7*(3), 105.

JAHROMI, A. T., STAKHOVYCH, S., & EWING, M. (2014). Managing B2B customer churn, retention and profitability. *Industrial Marketing Management*, *43*(7), 1258-1268.

JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R., & TAYLOR, J. (2023). *An Introduction to Statistical Learning : With Applications in Python*. Springer International Publishing. https://doi.org/10.1007/978-3-031-38747-0

KLESEL, M., & WITTMANN, H. F. (2025). Retrieval-Augmented Generation (RAG). *Business & Information Systems Engineering*, *67*(4), 551-561. https://doi.org/10.1007/s12599-025-00945-3

LEWIS, P., PEREZ, E., PIKTUS, A., PETRONI, F., KARPUKHIN, V., GOYAL, N., & KIELA, D. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks [arXiv :2005.11401]. https://doi.org/10.48550/arXiv.2005.11401

MA, Y. (2026). B2C e-commerce revenue in Brazil 2019-2025. https://www.statista.com/statistics/222115/online-retail-revenue-in-brazil-projection/

MANNING, C. D., RAGHAVAN, P., & SCHÜTZE, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

MANZOOR, A., QURESHI, M. A., KIDNEY, E., & LONGO, L. (2024). A review on machine learning methods for customer churn prediction and recommendations for business practitioners. *IEEE Access*, *12*, 70434-70463.

MAO, Y., LIU, Q., & ZHANG, Y. (2024). Sentiment analysis methods, applications, and challenges : A systematic literature review. *Journal of King Saud University-Computer and Information Sciences*, *36*(4), 102048.

MASYHURI, M. (2022). Key Drivers of Customer Satisfaction on the E-Commerce Business. *East Asian Journal of Multidisciplinary Research*, *1*(4), 657-670. https://doi.org/10.55927/eajmr.v1i4.405

MIKOLOV, T., CHEN, K., CORRADO, G., & DEAN, J. (2013). Efficient Estimation of Word Representations in Vector Space [arXiv :1301.3781]. https://doi.org/10.48550/arXiv.1301.3781

MORGAN, N. A. (2012). Marketing and business performance. *Journal of the Academy of Marketing Science*, *40*(1), 102-119. https://doi.org/10.1007/s11747-011-0279-9

MURPHY, K. P. (2012). *Machine Learning : A Probabilistic Perspective*. MIT Press.

NGUYEN, L., & CHANUT, O. (2018). Mesurer la qualité de service d'un site e-marchand : état de l'art des échelles de mesure et perspectives de recherche.

PANG, B., & LEE, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1-135.

REIMERS, N., & GUREVYCH, I. (2019). Sentence-BERT : Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3980-3990. https://doi.org/10.18653/v1/D19-1410

TABOADA, M., BROOKE, J., TOFILOSKI, M., VOLL, K., & STEDE, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, *37*(2), 267-307.

TIRUNILLAI, S., & TELLIS, G. J. (2014). Mining marketing meaning from online chatter : Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research*, *51*(4), 463-479.

VAN DOORN, J., & VERHOEF, P. C. (2008). Critical incidents and the impact of satisfaction on customer share. *Journal of Marketing*, *72*(4), 123-142.

VERBEKE, W., MARTENS, D., MUES, C., & BAESENS, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, *38*(3), 2354-2364.

ZHAO, W. X., ZHOU, K., LI, J., TANG, T., WANG, X., HOU, Y., & WEN, J.-R. (2023). A Survey of Large Language Models.

Annexes

1 Distribution des notes de review sur le jeu d'entraînement

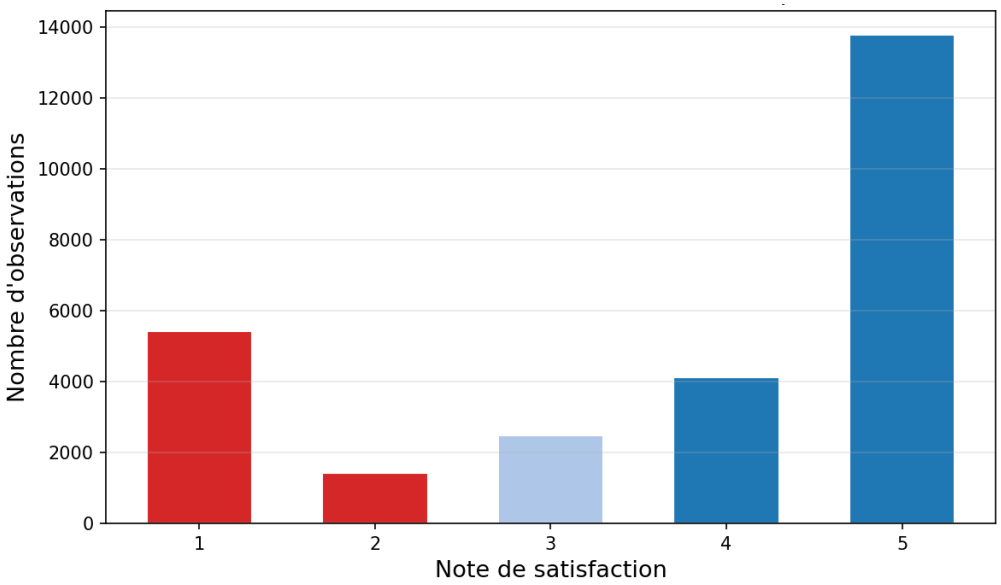


FIGURE 6 – Distribution des notes de review

2 Liste complète des features M1

TABLE 7 – Liste complète des features M1

Feature	Définition	Type	Famille
recency_days	Nombre de jours depuis la dernière commande	Numérique continu	RFM
frequency	Nombre de commandes (= 1 pour tous les clients)	Numérique entier	RFM
monetary	Valeur totale dépensée par le client	Numérique continu	RFM
n_items	Nombre d'articles dans la commande	Numérique entier	Comportementale
n_sellers	Nombre de vendeurs impliqués dans la commande	Numérique entier	Comportementale

Feature	Définition	Type	Famille
delivery_delay_days	Écart entre date de livraison estimée et réelle	Numérique continu	Comportementale
payment_value	Montant total du paiement	Numérique continu	Comportementale
is_late	Indicateur binaire de retard de livraison	Binaire (0/1)	Comportementale
processing_time_days	Délai entre la commande et l'expédition	Numérique continu	Comportementale
n_installments	Nombre de mensualités du paiement	Numérique entier	Comportementale
freight_ratio	Part des frais de port dans le montant total	Numérique continu	Comportementale
payment_boleto	Mode de paiement : boleto (one-hot)	Binaire (0/1)	Paiement
payment_credit_card	Mode de paiement : carte de crédit (one-hot)	Binaire (0/1)	Paiement
payment_debit_card	Mode de paiement : carte de débit (one-hot)	Binaire (0/1)	Paiement
payment_voucher	Mode de paiement : voucher (one-hot)	Binaire (0/1)	Paiement
state	État brésilien du client (target encoding)	Numérique encodé	Géographique
category	Catégorie principale du produit acheté (target encoding)	Numérique encodé	Produit
product_weight_g	Poids du produit en grammes	Numérique continu	Produit

3 Base de connaissances du pipeline RAG

Dimension : Livraison

ID	Source	Passage intégral
liv_01	Masyhuri (2022)	Fulfillment and reliability aim to fulfill customer orders correctly and deliver them on time. A negative outcome in fulfillment and reliability could affect overall customer satisfaction and ruin the future of e-commerce companies.
liv_02	Masyhuri (2022)	It is important for the company to provide service correctly by executing customer transactions correctly and delivering customer orders in a timely manner.
liv_03	Masyhuri (2022)	Order fulfillment is a key determinant of customer satisfaction. Fulfillment and reliability have a significant impact on e-customer satisfaction.
liv_04	Nguyen & Chanut (2018)	La fiabilité de la livraison et le traitement des commandes ont un impact significatif sur la qualité de service perçue. Les consonautes sont de plus en plus sensibles au service logistique.
liv_05	Nguyen & Chanut (2018)	Les délais de livraison, la traçabilité en temps réel et les modalités de retour constituent des facteurs déterminants de la qualité perçue du service électronique et de l'intention de réachat.
liv_06	Nguyen & Chanut (2018)	Questions clés pour le consonaute : la rapidité de la livraison, le respect du délai annoncé, la traçabilité des informations et les modalités de retour simplifiées.
liv_07	Nguyen & Chanut (2018)	Le service logistique est susceptible de créer un avantage concurrentiel pour les e-commerçants. La partie aval de l'achat reste un critère décisif d'évaluation.
liv_08	Tirunillai & Tellis (2014)	User-generated content allows firms to extract latent dimensions of customer satisfaction including objective dimensions such as delivery speed and reliability that dominate vertically differentiated markets.
liv_09	Archak et al. (2011)	Consumer reviews frequently mention shipping and delivery as distinct product features with associated opinions, reflecting customer satisfaction or dissatisfaction with the fulfillment process.

ID	Source	Passage intégral
liv_10	Masyhuri (2022)	Customers express dissatisfaction when orders arrive late, are damaged during shipping, or when tracking information is unavailable. Delivery failures directly reduce repurchase intentions.
liv_11	Nguyen & Chanut (2018)	Customers are satisfied when products arrive before or on the announced delivery date. Surpassing delivery expectations positively impacts overall service quality perception.

Dimension : Qualité produit

ID	Source	Passage intégral
qua_01	Masyhuri (2022)	Performance deterioration is a threat to customer loyalty. Even when a customer repurchases constantly, threats to loyalty arise from performance deterioration or product quality issues.
qua_02	Al-Ayed (2022)	When the performance of a product is lower than expected, negative disconfirmation occurs, causing dissatisfaction and reducing repurchase intentions.
qua_03	Al-Ayed (2022)	Positive or negative disconfirmation is formed when expectations before and after purchase are compared, which affects overall satisfaction with a product.
qua_04	De Caigny et al. (2020)	Product reviews are multifaceted. The textual content of reviews is an important determinant of consumer choices, over and above the valence and volume of reviews, because products have multiple attributes.
qua_05	De Caigny et al. (2020)	By compressing a complex review to a single number, we implicitly assume that product quality is one-dimensional, whereas products have multiple attributes and different attributes can have different levels of importance to consumers.

ID	Source	Passage intégral
qua_06	De Caigny et al. (2020)	Consumers evaluate product quality based on specific features such as picture quality, ease of use, battery life, or design. Negative evaluations of key product attributes directly reduce purchase likelihood.
qua_07	Tirunillai & Tellis (2014)	Quality is a multidimensional construct. The dimensions of quality are critical because they constitute the basis on which consumers evaluate brands and firms compete with one another.
qua_08	Tirunillai & Tellis (2014)	For vertically differentiated markets, objective dimensions dominate and are similar across markets. Customers explicitly mention product performance, durability, and conformity to description in reviews.
qua_09	Archak et al. (2011)	Each opinion phrase about a product feature represents a dimension of consumer experience. Negative opinion phrases about key product features are associated with lower customer satisfaction and reduced sales.
qua_10	Nguyen & Chanut (2018)	La qualité des contenus du site marchand, incluant l'exactitude et la complétude des informations produit, constitue une dimension récurrente de la qualité de service électronique.
qua_11	Masyhuri (2022)	Customers express strong dissatisfaction when products arrive broken, damaged, or do not match the description. Product non-conformity is a primary driver of negative reviews and churn.
qua_12	De Caigny et al. (2020)	Positive product quality perceptions arise when products meet or exceed customer expectations. Customers mention excellent quality, great design, and ease of use as key drivers of satisfaction.

Dimension : Service client

ID	Source	Passage intégral
ser_01	Masyhuri (2022)	Customer service aims to answer and solve customers requests and problems as quickly as possible. The lack of pleasant customer experiences and personalized advice discourages online shopping.
ser_02	Masyhuri (2022)	A company's prompt response to customer inquiries has a positive effect on satisfaction. Poor customer service results in immediate dissatisfaction and customer loss.
ser_03	Masyhuri (2022)	E-commerce firms should handle customers problems and complaints in a friendly and courteous manner. Customer service has a significant impact on online customer satisfaction.
ser_04	Al-Ayed (2022)	Care refers to the attention paid to the consumer before and after the act of purchase to maintain long-term relationships. It is a key determinant of e-customer loyalty.
ser_05	Al-Ayed (2022)	Contact interactivity is the degree to which a website facilitates interaction with customers. It is a determinant factor that influences the desire to return to the website.
ser_06	Nguyen & Chanut (2018)	La réactivité et le contact constituent des dimensions clés de la qualité de service électronique post-achat, notamment dans la gestion des réclamations et le traitement des retours.
ser_07	Tirunillai & Tellis (2014)	User-generated content captures customer satisfaction with service interactions. Customers explicitly mention support quality, response time, and problem resolution in reviews.
ser_08	De Caigny et al. (2020)	Consumer reviews mention service-related features such as customer support, return process, and after-sales assistance as distinct dimensions influencing overall satisfaction.
ser_09	Masyhuri (2022)	Customers become highly dissatisfied when companies fail to respond to complaints or inquiries. Lack of customer service response is a strong signal of churn intent.

ID	Source	Passage intégral
ser_10	Archak et al. (2011)	Opinion phrases about customer support and service interactions represent key dimensions in consumer reviews. Negative service opinions are strongly associated with product dissatisfaction.

Dimension : Prix

ID	Source	Passage intégral
pri_01	Masyhuri (2022)	More than 60% of customers worldwide cite the lowest price as their first reason for visiting an e-commerce website. However, quality of service is more important than low price for retention.
pri_02	Nguyen & Chanut (2018)	Le prix ne présente pas toujours l'élément le plus important pour l'intention de revisite. La prestation de service est plus importante que le bas prix pour fidéliser le client.
pri_03	Al-Ayed (2022)	The Internet has created hyper-competitivity where transaction costs are low and customers have great flexibility to move from one website to another, creating competition based on price.
pri_04	Al-Ayed (2022)	Switching barriers in e-commerce depend on the consumer's perception of money, time and effort required to change provider. Low switching costs make price a central factor in retention.
pri_05	De Caigny et al. (2020)	Consumers evaluate products based on multiple attributes including price-quality ratio. A product perceived as overpriced relative to its quality generates negative reviews and reduces repurchase intent.
pri_06	Tirunillai & Tellis (2014)	Price and value for money constitute explicit dimensions mentioned in consumer reviews. Customers express dissatisfaction when perceived value does not match the price paid.

ID	Source	Passage intégral
pri_07	Archak et al. (2011)	Product features have different pricing power. Consumer reviews reveal which product attributes justify the price paid and which attributes generate dissatisfaction relative to price expectations.
pri_08	Masyhuri (2022)	Customers express dissatisfaction when they feel the product or service does not offer sufficient value for the price paid. Poor price-quality ratio is a recurring theme in negative reviews.

4 Learning curve M2

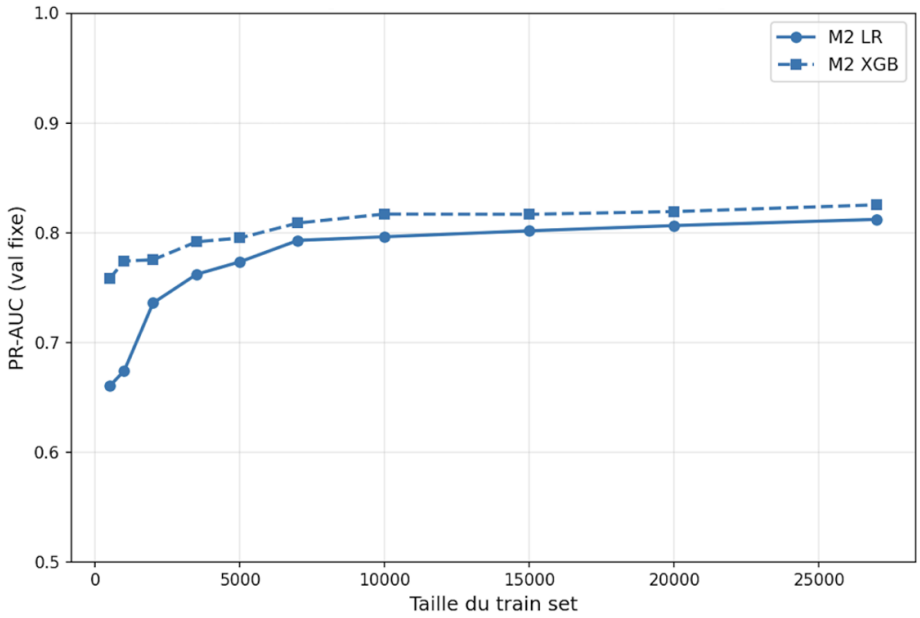


FIGURE 7 – Learning curve – Condition M2

5 Analyse de sensibilité *k* (FAISS)

TABLE 12 – Analyse de sensibilité *k* (FAISS)

k	**livraison (score)**	**qualité_produit (score)**	**service_client (score)**	**prix (score)**	**livraison (churn)**	**qualité_produit (churn)**	**service_client (churn)**	**prix (churn)**
1	0,628	0,504	0,432	0,084	−0,572	−0,434	−0,364	−0,024
2	0,638	0,471	0,403	0,135	−0,586	−0,396	−0,351	−0,082
3	0,599	0,456	0,326	0,084	−0,551	−0,388	−0,286	−0,024
5	0,465	0,273	0,346	−0,010	−0,442	−0,257	−0,331	+0,041
7	0,658	0,454	0,358	0,068	−0,601	−0,408	−0,315	0,000
10	0,669	0,460	0,410	0,068	−0,611	−0,396	−0,367	0,000

Note : corrélations de Pearson (scores RAG vs note de satisfaction) et point-bisériales (scores RAG vs churn), calculées sur 200 observations stratifiées du train pool.

6 Learning curve M3

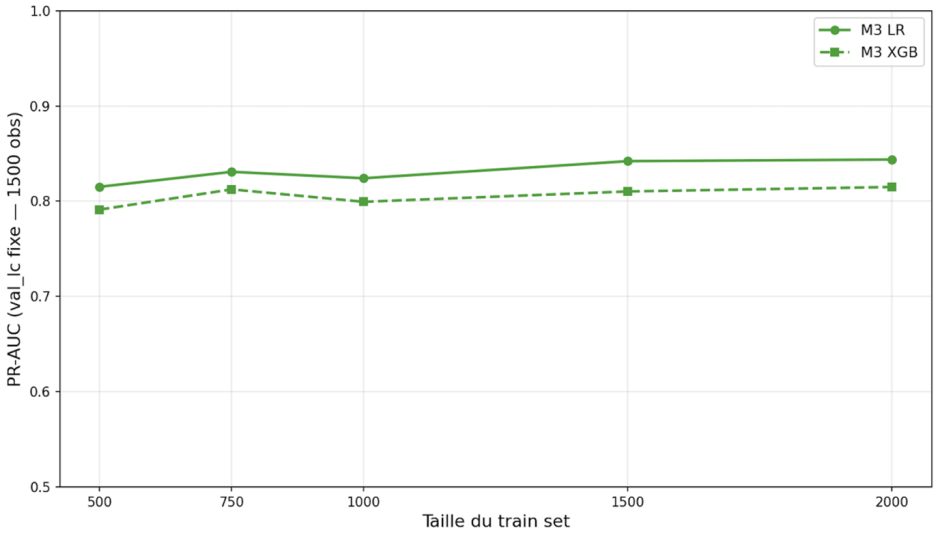


FIGURE 8 – Learning curve – Condition M3

7 Corrélations scores RAG vs note de satisfaction

TABLE 13 – Corrélations scores RAG vs note de satisfaction

Dimension	r	**p-value**
Livraison	0,646	$< 0{,}001$
Qualité produit	0,514	$< 0{,}001$
Service client	0,419	$< 0{,}001$
Prix	0,167	$< 0{,}001$

TABLE 14 – Moyenne des scores RAG par note de satisfaction

Note	**Livraison**	**Qualité produit**	**Service client**	**Prix**
1,0	0,184	0,365	0,340	0,478
2,0	0,208	0,309	0,439	0,483
3,0	0,425	0,440	0,479	0,498
4,0	0,616	0,627	0,498	0,503
5,0	0,746	0,726	0,535	0,516

8 Grilles d'hyperparamètres

Régression Logistique (toutes conditions)

TABLE 15 – Grille d'hyperparamètres – Régression Logistique

Hyperparamètre	**Valeurs testées**
C (régularisation L2)	1,0 ; 0,1 ; 0,01 ; 0,001
class_weight	balanced (fixe)
penalty	L2 (fixe)
max_iter	1 000 (fixe)

XGBoost M1 et M2

TABLE 16 – Grille d'hyperparamètres – XGBoost M1 et M2

Hyperparamètre	**Valeurs testées**
max_depth	3 ; 4
min_child_weight	5 ; 10
reg_lambda (L2)	2,0 ; 5,0
reg_alpha (L1)	0,1 ; 1,0
subsample	0,7 ; 0,8
colsample_bytree	0,7 ; 0,8
n_estimators	300 (fixe)
learning_rate	0,05 (fixe)
scale_pos_weight	ratio non-churners/churners (fixe)
eval_metric	aucpr (fixe)

XGBoost M3

TABLE 17 – Grille d'hyperparamètres – XGBoost M3

Hyperparamètre	**Valeurs testées**
max_depth	2 ; 3 ; 4
min_child_weight	5 ; 10 ; 15
reg_lambda (L2)	2,0 ; 5,0 ; 10,0
reg_alpha (L1)	0,1 ; 1,0 ; 2,0
subsample	0,6 ; 0,7 ; 0,8
colsample_bytree	0,6 ; 0,7 ; 0,8
n_estimators	300 (fixe)
learning_rate	0,05 (fixe)
scale_pos_weight	ratio non-churners/churners (fixe)
eval_metric	aucpr (fixe)

9 Code source

L'ensemble du code source développé dans le cadre de ce mémoire, du prétraitement des données à la génération des résultats, est disponible sur le dépôt GitHub suivant :

https://github.com/inassou25/nanat-inas-master-thesis

10 Résultats finaux complets sur le jeu de test

TABLE 18 – Résultats finaux complets sur le jeu de test

Méthode	**PR-AUC**	**ROC-AUC**	**F1**	**Recall**	**Precision**	**Seuil**
M1 LR	0,412	0,666	0,400	0,518	0,326	0,433
M1 XGB	0,441	0,675	0,412	0,413	0,412	0,504
M2 LR	0,829	0,961	0,801	0,887	0,730	0,627
M2 XGB	0,827	0,960	0,787	0,851	0,732	0,670
M3 LR	0,801	0,952	0,787	0,874	0,715	0,616
M3 XGB	0,821	0,955	0,808	0,889	0,741	0,629
M4 Zero-shot	0,816	0,967	0,830	0,925	0,754	0,800

11 Feature importance complète M1 XGB et M3 XGB

M1 XGB

TABLE 19 – Feature importance complète – M1 XGB

Feature	Importance
Retard de livraison (`is_late`)	0,648
Nombre d'articles (`n_items`)	0,133
Délai de retard en jours (`delivery_delay_days`)	0,078
Nombre de vendeurs (`n_sellers`)	0,040
Délai de traitement (`processing_time_days`)	0,017
Catégorie produit (`category_encoded`)	0,013
Récence en jours (`recency_days`)	0,009
Poids du produit (`product_weight_g`)	0,008
État client (`state_encoded`)	0,007
Paiement voucher	0,007
Montant payé (`payment_value`)	0,007
Montant total (`monetary`)	0,006
Ratio frais de port (`freight_ratio`)	0,006
Nombre de mensualités (`n_installments`)	0,006
Paiement boleto	0,005
Paiement carte de crédit	0,005
Paiement carte de débit	0,005

M3 XGB

TABLE 20 – Feature importance complète – M3 XGB

Feature	Importance
Score RAG — Livraison	0,253
Score RAG — Qualité produit	0,161
Score RAG — Service client	0,105
Retard de livraison (`is_late`)	0,102
Livraison mentionnée (binaire)	0,085
Délai de retard en jours (`delivery_delay_days`)	0,065
Nombre d'articles (`n_items`)	0,051
Catégorie produit (`category_encoded`)	0,031
État client (`state_encoded`)	0,018
Récence en jours (`recency_days`)	0,018
Paiement boleto	0,017
Poids du produit (`product_weight_g`)	0,017
Montant total (`monetary`)	0,015
Délai de traitement (`processing_time_days`)	0,015
Nombre de mensualités (`n_installments`)	0,014
Montant payé (`payment_value`)	0,014
Ratio frais de port (`freight_ratio`)	0,011
Qualité produit mentionnée (binaire)	0,009

12 Analyse de robustesse (Régression)

Résultat des métriques par condition

TABLE 21 – Métriques par condition – Analyse de robustesse

Méthode	**MAE**	**RMSE**	**Spearman**
M1 Ridge	1,170	1,394	0,234
M1 XGB	1,102	1,370	0,245
M2 Ridge	0,617	0,837	0,707
M2 XGB	0,592	0,824	0,709
M3 Ridge	0,684	0,883	0,684
M3 XGB	0,644	0,835	0,676
M4 Zero-shot	0,401	0,843	0,809

Tests de Wilcoxon – Significativité des différences

TABLE 22 – Tests de Wilcoxon – Significativité des différences

Comparaison	**p-value**
M1 XGB vs M2 Ridge	$< 0{,}001$
M1 XGB vs M3 XGB	$< 0{,}001$
M2 Ridge vs M3 XGB	$< 0{,}001$
M1 XGB vs M4 Zero-shot	$< 0{,}001$
M2 Ridge vs M4 Zero-shot	$< 0{,}001$

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management
Chaussée de Binche 151, 7000 Mons, Belgique | **www.uclouvain.be/lsm**