République du Sénégal

Un Peuple-Un But-Une Foi

* * * * * * * * *

Agence nationale de la Statistique et de la Démographie

* * * * * * * * *

Ecole nationale de la Statistique et de l'Analyse économique Pierre NDIAYE

* * * * * * * * *

Projet de Data Mining/Machine Learning

THEME

Prédiction de l'attrition des clients d'une Banque

Rédigé par :

SAMB Coumba

SOW Doulo

TRAORE Harouna

Elèves Ingénieurs des Travaux Statistiques

Sous la supervision de :

M. Abdou Aziz NDIAYE

Enseignant à l'ENSAE

Année scolaire 2020/2021

Prédiction de l'attrition des clients d'une Banque

Document rédigé par :
SAMB Coumba, SOW Doulo & TRAORE Harouna
Élèves Ingénieurs des Travaux Statistiques (ITS) en 4e année à l'ENSAE

Chargé du cours de Datamining :
Monsieur Abdou Aziz NDIAYE
Enseignant à l'ENSAE

Décharge

L'École nationale de la Statistique et de l'Analyse économique PIERRE NDIAYE (ENSAE) de Dakar n'entend donner aucune approbation, ni improbation aux idées émises dans ce rapport. Elles sont propres à leurs auteurs et doivent être considérées comme telles.

Sommaire

Décharge i

Sommaire iii

Liste des graphiques iv

Liste des tableaux v

Résumé vi

Preambule 2

Contextualisation 2

Introduction 3

I Etude préliminaire : Exploration et préparation des données 5

1 Présentation et traitement de la base de données 6

2 Profilage des clients et choix des variables 11

II Modélisation et présentation des résultats 17

3 Les différentes techniques de modélisation 18

4 Etude comparative des techniques 25

Postambule 33

Difficultés et Limites de l'étude 33

Conclusion 34

Références Bibliographiques i

Annexes **ii**

Table des matières **viii**

Liste des graphiques

2.1 *Répartition des clients selon l'attrition* 12
2.2 *Matrice de corrélation des variables* 14
2.3 *Importance des variables retenues* 16

4.1 *Courbe d'apprentissage du modèle XGBoost* 30
4.2 *Courbe de ROC du modèle XGBoost* 31

Liste des tableaux

1.1 *Les variables socio-démographiques* 7
1.2 *Les variables relatives à la carte bancaire* 7
1.3 *Variables relatives à la relation avec la banque* 8
1.4 *Les variables relatives aux transactions* 8

2.1 *Les variables retenues pour le modèle* 15

4.1 *Résultats sur les réseaux de neurones* 27
4.2 *Résultats sur les modèles ensemblistes* 28
4.3 *Résultats sur les modèles standards* 29
4.4 *Les résultats sur les modèles sans surapprentissage* 30
5 *L'attrition selon le sexe des clients* iv
6 *L'attrition selon la situation matrimoniale* iv
7 *L'attrition selon le type de la carte* iv
8 *L'attrition selon la catégorie de revenu* v
9 *L'attrition selon le niveau d'éducation* v
10 *Les variables quantitatives et leurs satatistiques* vi

Résumé

Ce projet entre dans le cadre du cours de DatMining effectué en dernière année dans la filière ITS, particulièrement pour l'option Statistique décisionnelle. Il se fait en guise d'évaluation des compétences et des connaissances acquises au cours. L'objectif est de mettre en place un modèle capable de prédir l'attrition des clients d'une Banque.
Aprés une analyse et un processing sur les données afin de préparer nos données à la modélisation, une brève description des variables disponibles et des clients selon plusieurs caractéristiques (variables socio-démographiques, variables relatives aux aux activités des clients au niveau de la Banque) a été effectuéé. Globalement, nous avons 1627 cas d'attrition (soit 16,10%) contre 8500 (soit 83,90%).
Au total, nous avons eu à tester **neuf (9) modèles** (Modèle Logit, SVM, arbre de décision, réseaux de neurones, XGBoost, AdaBoost, KNN, Random Forest, et Bagging). Aprés avoir au préalable effectué une sélection des variables les plus discriminantes au regard de notre variable cible, nous sommes passés à l'implémentation des modèles précités.
L'algorithme SelectKbest nous a permis de retenir 5 variables (nombre de mois d'inactivité au cours des 12 derniers mois, Limite de crédit sur la carte de crédit, Montant total de la transaction (12 derniers mois), Nombre total de transactions (12 derniers mois) et Changement du nombre de transactions.). Notre objectif global étant de pouvoir prédir si un client quitte ou non la Banque, nous devons donc minimiser le nombre de faux négatifs (dire qu'un client reste dans la banque alors qu'il quitte). Nous précisons, ici, que Un positif est celui qui quitte (variable cible = 1). L'indicateur permettant de prendre en compte cet objectif est donc la **sensibilité** [1] qu'il faut maximisé. On a cependant tenu en compte d'autres indicateurs (précision, F1-score ...) et avons retenu le XGBoost qui trouve le meilleur compromis parmi les modèles qui ne surapprennent pas.

1. $Sensitivity = \frac{VP}{VP+FN}$

PREAMBULE

Contexte d'exécution du projet

Basée à Dakar au Sénégal, l'ENSAE fait partie intégrante de l'Agence Nationale de la Statistique et de la Démographie (ANSD). Elle est une école panafricaine d'enseignement supérieur en statistique et est membre à part entière du réseau des trois (03) Grandes Écoles Africaines de Statistiques tout comme l'École Nationale Supérieure de Statistique et d'Économie Appliquée (ENSEA) d'Abidjan et l'Institut Sous régional de la Statistique et de l'Économie Appliquée (ISSEA) de Yaoundé. Ainsi, l'ENSAE forme des cadres supérieurs et performants dans trois filières à savoir :

- Techniciens Supérieurs de la Statistique (TSS) ;
- Ingénieurs des Travaux Statistiques (ITS) ;
- Ingénieurs Statisticiens Economistes (ISE).

Les étudiants de la filière ITS sont appelés à effectuer une formation d'une durée de quatre années (4) où ils sont appelés à maitriser des concepts mathématiques, en particulier statistiques combinées à l'Economie et à l'Informatique. A partir de la dernière année, deux vois se dégagent : Finance et Statistique décisionnelle. Etant dans ce dernier cas, nous sommes appelés à faire un cours de Datamining censé nous aider à maitriser les téchniques de Machine Learning et tant d'autres. C'est dans le cadre de ce cours que s'incrit ce document relatif à son évaluation en guise de projet. Ce projet, nous met dans une situation réelle d'application des techniques du DataMining. Son objectif principal est de mettre en place un modèle de **PREDICTION DE L'ATTRITION DES CLIENTS D'UNE BANQUE.**

Introduction

Problématique

Dans un contexte économique et financier de plus en plus décloisonné, causé par la mondialisation, les acteurs bancaires sont livrés à une compétition accrue pour se maintenir sur le marché financier. Il incombe alors à chaque responsable bancaire d'assurer sa pérennité en mettant en place des politiques de maintien de sa clientèle en trouvant des moyens de limiter l'attrition des clients. Car, dans un système concurrentiel caractérisé par une récurrence et une alternance des faillites, la prévention des risques et une bonne capacité de rétention de sa clientèle sont de mise. Il faudrait donc, non seulement assurer un bon fonctionnement interne des banques et limiter le risque de non-remboursement des clients ou de survenance d'un sinistre, mais adopter des méthodes permettant d'atteindre sa clientèle et d'anticiper sa demande pour la fidéliser. Comment, cependant, arriver à cette fin ?
Ainsi, à travers son historique, la banques peut jauger le comportement de ses clients et prévoir les risques d'attrition. Pour bien mener cette politique de fidélisation les banques cherchent au tant que faire se peut d'avoir le maximum d'informations sur leurs clients et cerner leurs comportements afin d'adapter l'offre de service. Les banques collectent donc régulièrement des données clients et tentent de les valoriser. Pour atteindre ces objectifs, la disponibilité des données de la clientèle et l'utilisation des outils statistiques, notamment la méthode scoring, sont essentielles

Objectif

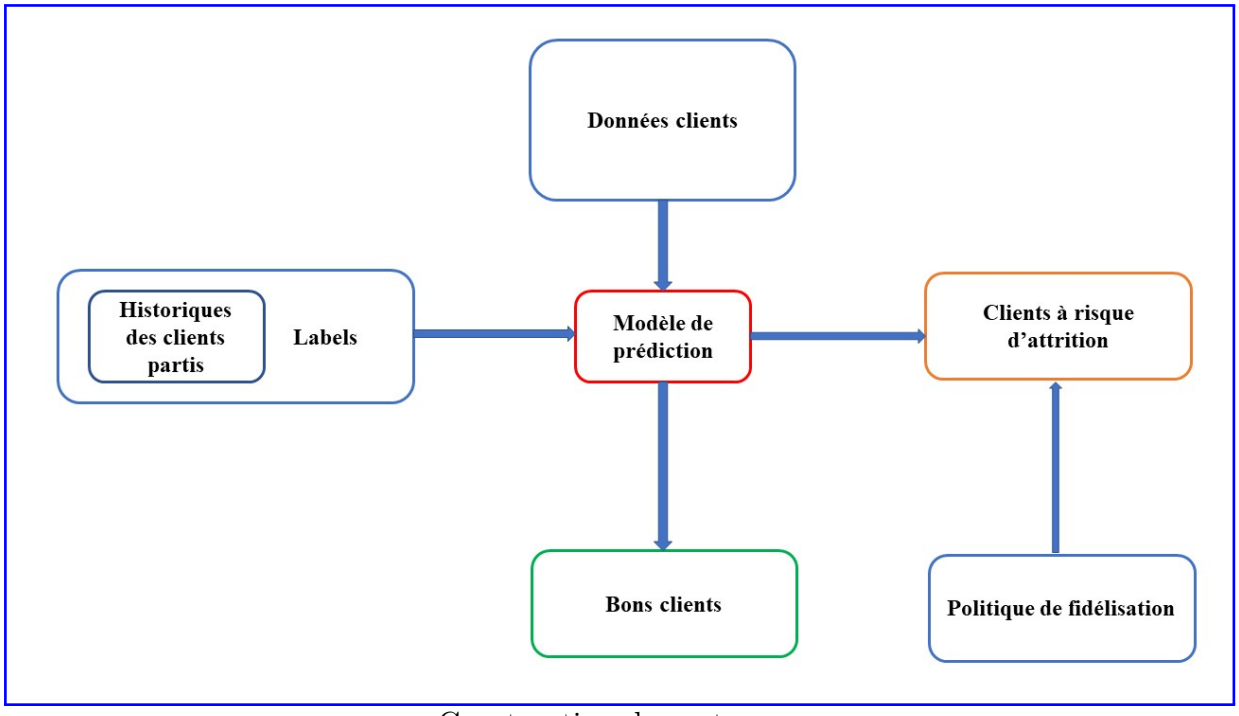
Le scoring client est une méthode statistique permettant de classer les clients, en attribuant un score à chacun d'eux, au regard de certaines caractéristiques jugées liées au phénomène ou évènement à prédire. Les scores sont généralement de cinq types : (1) le score d'octroi ou d'acceptation qui est calculé pour un nouveau client ou un client à faible activité avec la banque ; (2) le score de comportement ou de risque qui est la probabilité pour un client de rencontrer un incident de paiement ou de remboursement ; (3) le score d'appétence, qui revêt de deux types : le score de propension à consommer et le score d'affinité. (4) Le score de recouvrement qui évalue le montant susceptible d'être récupéré sur un compte en cas de contentieux. (5) Le score d'attrition qui évalue la probabilité de quitter la banque. Ce dernier type fait l'objet de ce document et traite spécifiquement de **la mise en œuvre d'un modèle de prédiction d'attrition de clients**, à l'aide des

techniques du DataMining, à partir de données bancaires.

Démarche et méthodologie

Pour mener à bien cette étude de ciblage de la population à risque d'attrition, nous adoptons la démarche suivante : (1)Une exploration et présentation des données fera l'objet de la première partie qui traitera spécifiquement de la présentation et du nettoyage de la base, du profilage des clients, à l'aide des techniques d'analyse univariée et multivariée, et de méthodes de choix des différentes variables. (2) La seconde partie, composée de deux chapitres, traite globalement de la modélisation et présente les résultats obtenus. Le chapitre premier de cette partie, donne les différentes techniques de modélisation que nous testerons et le second passe à l'étude comparatif afin d'en retenir une. (3) En définitif, le postambule se penche sur les difficultés et limites de l'étude avant de finir avec les blians relatfs à l'étude et au cours de DataMining.

En résumé le travail à faire se résume par ce graphique :



Construction des auteurs

Première partie

Etude préliminaire : Exploration et préparation des données

Chapitre

1

Présentation et traitement de la base de données

Dans ce chapitre, il s'agira de faire un descriptif de la base en première lieu avant de passer à son traitement préliminaire. Ainsi, dans la première section, nous décrirons la base selon ses dimensions, regrouperons les variables en catégories avant de passer à leur descriptif standard. Dans la seconde section, nous passerons au traitement des données notamment des valeurs manquantes, des doublons, des valeurs extrêmes avant de terminer avce la discrétisation et la catégorisation des variables.

1.1 Description de la base de données 6
1.1.1 Variables socio-démographiques (6) 6
1.1.2 Variables bancaires (13) 7
1.2 Traitement préliminaire des données 8
1.2.1 Valeurs manquantes 9
1.2.2 Doublons 9
1.2.3 Valeurs extrêmes et valeurs aberrantes 9
1.2.4 Discrétisation 10
1.2.5 Catégorisation 10

1.1 Description de la base de données

La base de données qui nous est soumise est composée de dix mille cent vingt-sept (10127) individus et vingt-et-une (21) variables dont six (06) catégorielles ou qualitatives et quinze (15) de type quantitatif. Pour une bonne interprétation, ces variables peuvent être regroupés de la manière suivante :

1.1.1 Variables socio-démographiques (6)

Ces variables caractérisent le client lui-même et portent notamment sur le sexe, l'âge la situation matrimoniale, le niveau d'éducation, le nombre de personne que la personne

à en charge, et le revenu du client. Le tableau suivant donne le descriptif brief de ces variables.

Tableau 1.1 – *Les variables socio-démographiques*

Les variables socio-démographiques	
Noms dans la base	**Description**
Customer_Age	Âge du client en années
Gender	Sexe du client (M = Homme, F = Femme)
Dependent_count	Nombre de personnes à charge
Education_Level	Niveau d'éducation
Income_Category	Catégorie de revenu annuel
Marital_Status	situation maritale

Source : Base de données

1.1.2 Variables bancaires (13)

Ces variables mettent en exergue les caractéristiques du client au regard de sa relation avec la banque (Période de relation avec la banque, ligne de crédit. . .), de sa carte de crédit (type de carte, limite de crédit sur la carte . . .) et des différentes activités effectuées au sein de la banque (transactions). De ce fait, nous avons encore décidé de regrouper ces variables dans les sous-groupes suivants :

Carte bancaire (4)

Ces variables caractérisent l'individu selon la carte. Le tableau suivant donne un aperçu de ces variables. Leurs caractéristiques et quelques statistiques descriptives pour voir de manière globale le profil de la population pour ces types de variables, sont renseignées en annexe.

Tableau 1.2 – *Les variables relatives à la carte bancaire*

Les variables relatives à la carte bancaire	
Noms dans la base	**Description**
Card_Category	Type de carte
Credit_Limit	Limite de crédit sur la carte de crédit
Avg_Utilization_Ratio	Taux d'utilisation moyen de la carte
Total_Revolving_Bal	Solde renouvelable total sur la carte de crédit

Source : Base de données

Relation bancaire (5)

Dans ce groupe, nous mettons les variables qui décrivent le client selon la relation entretenue avec la banque. Le tableau suivant donne un résumé de ces variables :

Tableau 1.3 – *Variables relatives à la relation avec la banque*

Variables relatives à la relation bancaire	
Noms dans la base	**Description**
Months_on_book	Période de relation avec la banque
Total_Relationship_Count	Nombre total de produits détenus par le client
Months_Inactive_12_mon	Nombre de mois d'inactivité au cours des 12 derniers mois
Contacts_Count_12_mon	Nombre de contacts au cours des 12 derniers mois
Avg_Open_To_Buy	Ligne de crédit ouverte à l'achat (moyenne des 12 derniers mois)

Source : Base de données, nos calculs.

Transactions (4)

Ces variables caractérisent l'individu selon les transactions effectuées au niveau de la Banque. Le tableau suivant donne un aperçu de ces variables. Leurs caractéristiques et quelques statistiques descriptives pour voir de manière globale le profil de la population pour ces types de variables, sont renseignées en annexe.

Tableau 1.4 – *Les variables relatives aux transactions*

Les variables relatives aux transactions	
Noms dans la base	**Description**
Total_Amt_Chng_Q4_Q1	Changement du montant de la transaction (T4 par rapport au T1)
Total_Trans_Amt	Montant total de la transaction (12 derniers mois)
Total_Trans_Ct	Nombre total de transactions (12 derniers mois)
Total_Ct_Chng_Q4_Q1	Changement du nombre de transactions (T4 par rapport au T1)

Source : Base de données

1.2 Traitement préliminaire des données

Cette section est réservée au traitement de nos données. Il s'agira de traiter les données manquantes (si elles sont présentes), de corriger les valeurs rares ou extrêmes qui peuvent

déséquilibrer le modèle, de traiter les valeurs aberrantes qui conduisent à un faux modèle (car basé sur des mesures fausses). En outre dans l'utilisation éventuelle des méthodes à hypothèses de normalités (comme l'analyse discriminante de Ficher), nous effectuerons des tests de normalité, d'homoscédasticité pour s'assurer de la validité des modèles.

1.2.1 Valeurs manquantes

Une valeur manquante, pour une variable donnée, correspond à une donnée non renseignée et qui devait l'être en réalité. Après inspection de notre base de données, il ressort que la totalité des variables de la base sont complètes et donc sans aucune valeur manquante. Dans la suite, il s'agira alors de faire un processing de ces valeurs renseignées afin de détecter quelques anomalies.

1.2.2 Doublons

On parle de doublons quand une ligne se répète au moins deux fois. En d'autres termes, un même individu figure deux fois dans la base. Pour pallier ces difficultés qui peuvent impacter sur les résultats car le point de cet individu devient élevé, il faut passer à une suppression de ces lignes qui s'ajoutent. Dans notre base de données, cependant, après manipulation, nous notons que toute les lignes sont uniques et donc aucune donnée ne se répète.

1.2.3 Valeurs extrêmes et valeurs aberrantes

Une valeur aberrante est une valeur qui diffère de façon significative de la tendance globale des autres observations quand on observe un ensemble de données ayant des caractéristiques communes. La valeur aberrante est fausse et correspond à une valeur impossible, une erreur de mesure, une erreur de calcul ou fausse déclaration sur la variable. A côté, nous avons les valeurs extrêmes ou atypiques. Ces valeurs non faussent caractérisent un individu qui s'écarte largement de la population au regard de cette variable. La valeur atypique peut être à gauche (donc très petite par rapport à l'ensemble) ou à droite (donc très grande par rapport à l'ensemble des autres valeurs).
Une phase cruciale dans le traitement est la détection de ces valeurs qui peuvent déséquilibrer notre modèle et induire à une mauvaise prédication (car le modèle reposera sur des valeurs assez rares et donc difficiles à trouver, en valeur d'entrée des caractéristiques d'un individu, pour une phase ou de prédiction). Pour détecter ces valeurs, plusieurs méthodes sont possibles. Globalement, nous avons deux types : les méthodes graphiques et les méthodes statistiques. Les premières consistent à observer la variable à travers une visualisation, soit du boxplot, soit du nuage de points, soit de l'histogramme... Les dernières concernent les tris à plat, les contrôles logiques et les tests statistiques. Comme

tests statistiques, nous avons le test de GRUBBS, le test de DIXON, le test de HAMPLE. Dans notre étude, nous avons adopté la méthode des boxplots et celle de GRUBBS [1] . Une fois détectées, nous sommes passés à la correction de ces valeurs dites aberrantes ou atypiques. Pour traiter ces valeurs, plusieurs méthodes s'offrent au statisticien. Nous avons plusieurs alternatives regroupées en sous-groupes : (1) rejet ou acceptation systématique et (2) traitement ou remplacement. La dernière sera retenue du au fait que le test de GRUBBS n'a pas détecter un nombre assez élevé de valeurs aberrantes pouvant affecter de manière significative notre base de données.

1.2.4 Discrétisation

Face à des variables continues, il est très utile parfois avant l'utilisation de ces variables de voir s'il est pertinent ou non de les mettre dans un regroupement ou classe. La discrétisation, bien que n'étant pas pertinente pour les arbres de décision, car ceux-ci l'intégrant automatiquement, lors d'une régression logistique, elle est souvent un moyen d'optimisation et d'augmentation du pouvoir prédictif.
Cependant, quel est le nombre de classes à retenir ? Comment choisir les bornes de ces classes ? Deux méthodes sont généralement utilisées : la discrétisation non supervisée et la discrétisation supervisée. La première se base sur les caractéristiques de la variable à discrétiser (quartiles, écart à la moyenne, classification) pour effectuer le découpage. La deuxième méthode, celle qui est retenue, s'inscrit dans un cadre prédictif et est guidée par une variable supposée à expliquer. Pour appliquer cette dernière méthode, le découpage se fera au regard de la variable à expliquer « Attrition_Flag ». Il existe, cependant, deux approches pour cette méthode : l'approche ascendante (comme le Chi-Merge de Kerber (1992) et l'algorithme de Fisher)) et l'approche descendante. Pour cette dernière approche, nous avons plusieurs variétés d'algorithmes dont la référence est l'algorithme MDLPC de Fayyad & Irani (1993). Cet dernier a été en premier lieu retenu, du fait de son adéquation à notre étude, mais un problème d'exécution de l'algorithme s'est posé. C'est ainsi, que la méthode automatique de fréquence égale a été retenue et a permis de discrétiser les variables quantitatives (âge, ancienneté, transactions. . .).

1.2.5 Catégorisation

Les algorithmes implantés pour les différentes méthodes que nous utiliserons nécessitent une catégorisation. Alors, nous avons, en premier lieu, sélectionner l'ensemble des variables catégorielles avant de passer à la discrétisation. Cette étape consiste à codifier les valeurs numériques au lieu de leur utilisation avec les labels des modalités.

1. Voir description en annexe

Chapitre

2

Profilage des clients et choix des variables

Dans ce chapitre, l'objectif principal est de comprendre nos données dans leur globalité. En d'autres termes, il s'agira de connaitre les grandes tendances de chaque variable prise individuellement, de voir les associations existantes entre variables c'est-à-dire jauger les rapports d'affinité entre nos variables. Cela permettra par la suite de pouvoir comprendre la structuration de notre population, les différentes classes qui se dégagent et d'avoir un résumé de variables (indicateurs) permettant de réduire nos dimensions, et indirectement nos paramètres à estimer, pour éviter un sur-apprentissage de notre modèle final. Ainsi, dans la suite après avoir décrit chaque variable, nous utiliserons les méthodes descriptives du DataMining afin de décrire globalement le comportement et les caractéristiques les clients de la banque.

2.1 Analyse univariée 11
2.1.1 Attrition des clients 11
2.1.2 Caractéristiques socio-professionnelles des clients 12
2.1.3 Relation des clients avec la banque 12
2.2 Analyse multivariée 13
2.3 Détection des variables les plus discriminantes 13
2.3.1 Présentation des critères de choix 14
2.3.2 Présentation des variables retenues 15

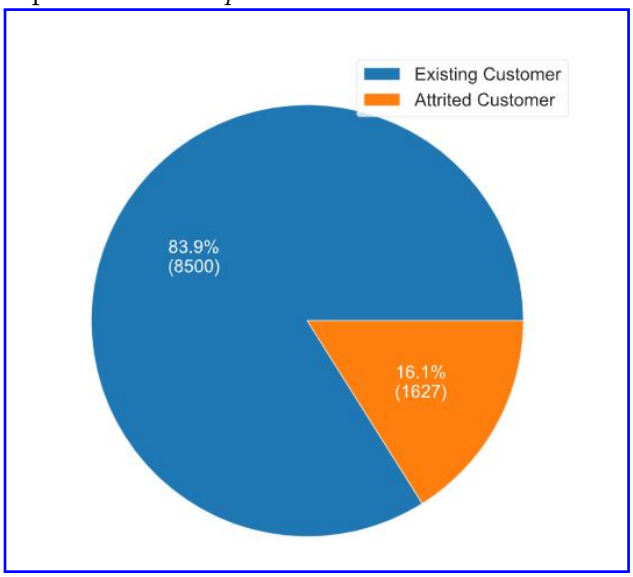
2.1 Analyse univariée

Cette partie se concentre sur les variables prises individuellement et met en exergue le profil des clients.

2.1.1 Attrition des clients

Il ressort de l'analyse du graphe ci-dessous que 83,9% (soit 8500 clients) des clients de la banque n'ont pas fait l'objet de départ contre 16,1% (soit 1627) pour ceux ayant quitté la banque.

Graphique 2.1 – *Répartition des clients selon l'attrition*



Source : Base de données, nos calculs.

2.1.2 Caractéristiques socio-professionnelles des clients

Profil des clients selon les caractériques socio-démographiques (les tableaux statistiques sont en annexe)

Il ressort des informations obtenues de la base de données que les clients sont en moyenne âgés de 46 ans et majoritairement constitués de femmes (53,58% contre 47,1%). Par rapport au niveau d'éducation, il est constaté que la majorité des clients sont du niveau supérieur (30,9%) suivie des individus de niveau Lycée (19,9%). Les doctorants représentent 4,5% des clients. S'agissant de la situation matrimoniale, les clients sont en majorité (46,3%) des mariés contre 38,9% de célibataires. Par rapport à la situation financière, il est à noter que 35,2% des clients ont un revenu inférieur à 40 milles dollars ($40K) contre 7,20% ayant un revenu de 120 milles dollars ($120K). Concernant les charges, en moyenne chaque client à en sa charge deux (2) personnes avec un maximum de 5 personnes.

2.1.3 Relation des clients avec la banque

Ici, nous ferons l'analyse selon les trois catégories de variables détectées plus haut (carte bancaire, relation avec la banque et transactions).

Description des individus (les tableaux statistiques sont en annexe)

Par rapport aux cartes bancaires, l'information principale à retenir est que plus de la moitié des clients de notre base de données possèdent une carte de type Blue (92,2%).

Très peu de clients détiennent des cartes des types Gold ou Platinum (moins de 1 % des clients). Nous notons que 75% des clients ont un taux d'utilisation de leurs cartes en deçà de 0,50%. La limite de crédit sur la carte de crédit s'établit, en moyenne, à 8600 (unités monétaires) avec un maximum de 34516 (unités monétaires). En moyenne, les clients ont un solde renouvelable total sur la carte de crédit d'environ 1160 (unités monétaires) avec un minimum et un maximum respectivement de 0 (unités monétaires) et 2517 (unités monétaires).
Concernant la relation avec la banque, il est à noter que tous les clients ont au moins eu un an de relation avec la banque (minimum 13 mois). En moyenne, la période de la relation avec la banque est de 35 mois soit environs 3 ans. Chaque client détient au moins un produit. Avec une moyenne du nombre de produits détenus qui est d'environ quatre (4), la médiane s'établit elle aussi à quatre (4). Ce qui veut dire qu'au moins 50% des clients détiennent plus de quatre (4) produits au niveau de la banque. Au niveau de l'activité, on peut dire que les clients sont peu actifs dans la mesure où près de 75% d'eux ont une période de non-activité inférieur à 3 mois au cours des douze derniers mois. En outre, 50% des clients ont eu au plus 2 mois de contacts avec la banque durant les douze derniers mois.
S'agissant des transactions, les clients ont effectué, en moyenne, 65 transactions avec un minimum de 10 et un maximum de 139.

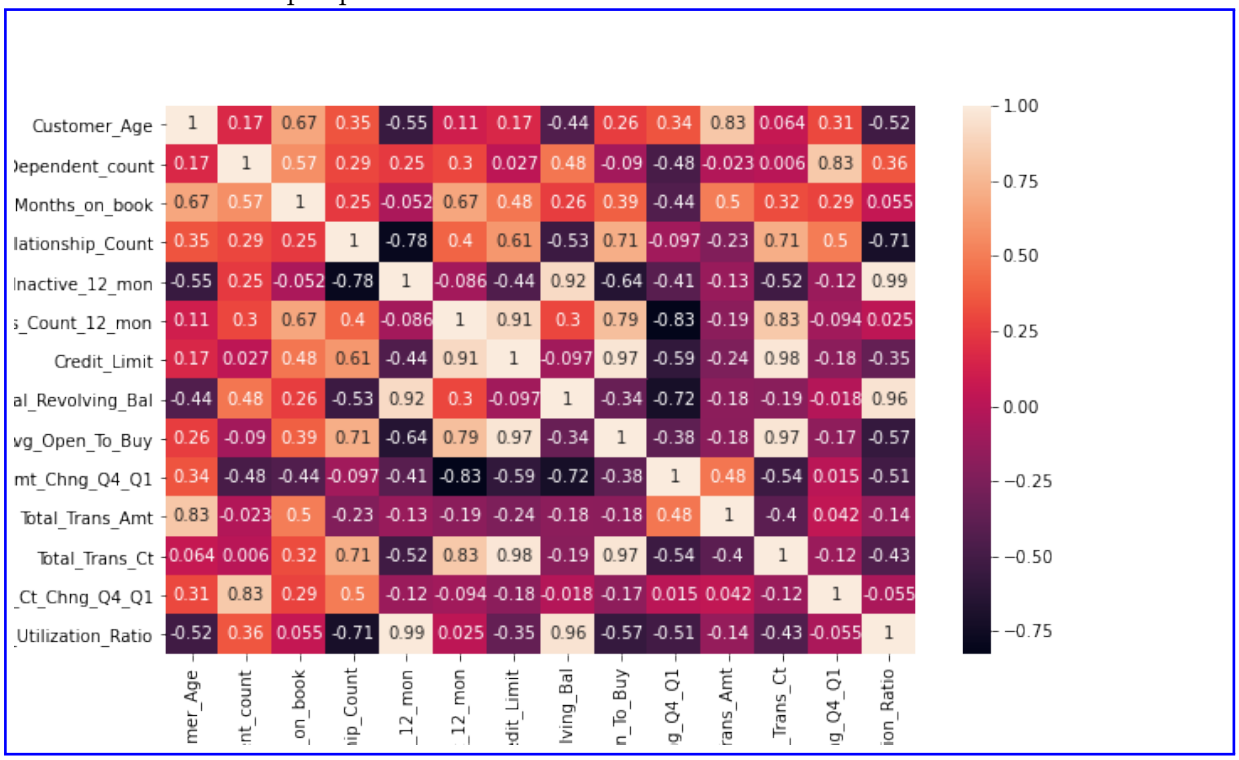
2.2 Analyse multivariée

Pour cette partie, nous avons calculé la matrice de corrélation permettant de mettre en exergue les liaisons existants entre nos différentes variables. Car ces informations peuvent permettre de bien choisir les variables à inclure dans nos modèles en complétant cette analyse avec une analyse multidimensionnelle. Cependant, dû au fait que nous avons décidé de ne retenir que la méthode SelectKbest, cette description multidimensionnelle ne sera qu'a but descriptif de la population et non du choix des variables.

2.3 Détection des variables les plus discriminantes

Lors d'une modélisation dont l'objectif est prédictif, une des étapes cruciales dans le preprocessing des variables explicatives est de voir celles qui sont les plus discriminantes. Dans le cas d'une prédiction avec les arbres de décision, cette étape n'est pas nécessaire car les algorithmes des arbres détectent eux-mêmes les variables les plus discriminantes et leur résultat n'est pas affecté par celles non-discriminantes. Dans la mesure où nous testons plusieurs algorithmes pour entrainer notre modèle, nous avons choisis de retenir une méthode de sélection des variables les plus discriminantes parmi les possibilités offertes par la littérature.

Graphique 2.2 – *Matrice de corrélation des variables*



Source : Base de données, nos calculs.

2.3.1 Présentation des critères de choix

L'étape de selection des variables à consister à choisir un sous ensemble de k variables parmi les n variables initiales à même d'expliquer l'attrition. La méthode utilisée est la selection univariée de variables (SelectKbest). Il s'agit d'un transformer utilisant les tests de dépendance (chi2, Information mutuelle, F-mesure) pour sélectionner un nombre K de variables ayant un lien fort avec la variable d'interêt. Selon les types de variables, l'algorithme utilise différentes mesures.

Cas des variables qualitatives

Pour le cas des variables qualitatives, l'algorithme essais de calculer la liaison qui existent entre la variable qualitative explicative avec la variable à prédire. La liaison est calculée avec la distance du Chi-2, en calculant la probabilité associée au V de Cramer.

Cas des variables quantitatives

S'agissant des variables quantitatives, la sélection se fait à l'aide de l'application d'un test paramétrique de la variance ANOVA ou un test non-paramétrique. Selon que les hypothèses de normalité et d'homoscédasticité sont respectées ou non, on applique res-

pectivement un test ANOVA ou non-paramétrique. Pour les détails dans l'algoriyhme, nous renvoyons le lecteur au livre renseigné en annexe (Stéphane TUFFERY à la page 63).

2.3.2 Présentation des variables retenues

L'implémentation a permis de retenir les variables suivantes :

Tableau 2.1 – *Les variables retenues pour le modèle*

Variables retenues après le le SelectKbest	
Noms dans la base	**Description**
Months_Inactive_12_mon (F0)	Nombre de mois d'inactivité au cours des 12 derniers mois
Credit_Limit(F1)	Limite de crédit sur la carte de crédit
Total_Trans_Amt(F2)	Montant total de la transaction (12 derniers mois)
Total_Trans_Ct(F3)	Nombre total de transactions (12 derniers mois)
Total_Ct_Chng_Q4_Q1(F4)	Changement du nombre de transactions (T4 par rapport au T1)

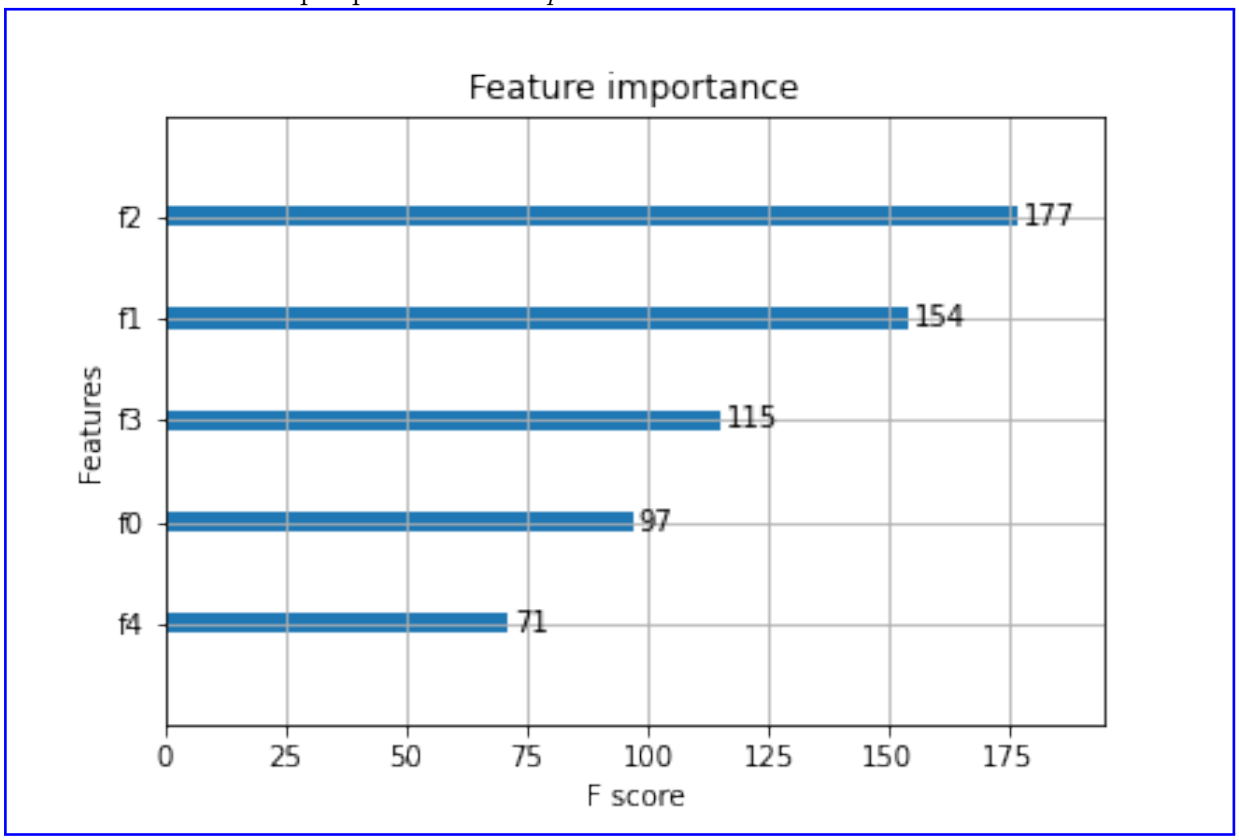
Source : Base de données, nos calculs.

La pertinence des variables choisies sont les suivantes :

- Un compte longtemps inactif les derniers pressages d'un abandon du service ;
- Moins le client détient de crédit dans son compte plus il est susceptible de quitter car le départ n'a pas trop de conséquence financière ;
- Un compte qui ne connait pas trop de mouvement (faible nombre et faible montant des transactions) laisse deviner un départ imminent.

Par suite, il est utile de déterminer la contribution relative de chaque variable dans la construction des arbres dans le modèle retenu à savoir le XGboost. En général, l'importance fournit un score qui indique l'utilité ou la valeur de chaque variable dans la construction des arbres de décision améliorés dans le modèle. Plus une variable est utilisée pour prendre des décisions dans les arbres de décision, plus son importance relative est élevée. L'importance est calculée pour un arbre de décision unique par le degré d'amélioration de la mesure de performance apporté par chaque point d'une variable, pondérée par le nombre d'observations dont le nœud est responsable. La mesure de la performance peut être la pureté (indice de Gini) utilisée pour sélectionner les points de partage ou une autre fonction d'erreur plus spécifique. Les importances des fonctionnalités sont ensuite moyennées sur tous les arbres de décision du modèle.
Le graphique ci-après indique l'importance des variables :

Graphique 2.3 – *Importance des variables retenues*



Source : Base de données, nos calculs.

Deuxième partie

Modélisation et présentation des résultats

Chapitre

3

Les différentes techniques de modélisation

L'objectif principal de ce chapitre est de présenter les différentes techniques de modélisation en Datamining censées répondre à la problématique de notre étude. Pour ce faire, nous passerons en revue les différentes téchniques du Datamining du point de vue théorique et de leur mise en œuvre.

3.1 Le modèle logistique 18
3.2 Le modèle Bagging 19
3.3 Le modèle XGboost 19
3.4 Le Réseau de neurone 20
3.5 Les arbres de décision 21
3.6 Le modèle Random Forest 22
3.7 Le modèle KNN 23
3.8 Le modèle AdaBoost 24
3.9 Le modèle SVM 24

3.1 Le modèle logistique

La régression logistique est une technique prédictive. Elle vise à construire un modèle permettant de prédire / expliquer les valeurs prises par une variable cible qualitative (le plus souvent binaire, on parle alors de régression logistique binaire ; si elle possède plus de 2 modalités, on parle de régression logistique polytomique) à partir d'un ensemble de variables explicatives quantitatives ou qualitatives, pour lequel cas, un codage est nécessaire. Le recours à ce modèle se justifie par le fait que la variable Y à expliquer prend les valeurs Y = 0 ou 1, ce qui rend impossible l'ajustement par la méthode des moindres carrés lorsque nous avons plus de deux individus. La fonction de densité du modèle est donnée par :

$$f(x) = \frac{exp(x)}{1 + exp(x)}$$

3.2 Le modèle Bagging

Le mot Bagging est une contraction de Bootstrap Aggregation. Le bagging est une technique utilisée pour améliorer la classification notamment celle des arbres de décision, considérés comme des « classifieurs faibles », c'est-à-dire à peine plus efficaces qu'une classification aléatoire.
En général, le bagging a pour but de réduire la variance de l'estimateur, en d'autres termes de corriger l'instabilité des arbres de décision (le fait que de petites modifications dans l'ensemble d'apprentissage entraînent des arbres très différents). Pour ce faire, le principe du bootstrap est de créer de « nouveaux échantillons » par tirage au hasard dans l'ancien échantillon, avec remise. L'algorithme, par exemple l'arbre de décision, est entraîné sur ces sous-ensembles de données. Les estimateurs ainsi obtenus sont moyennés (lorsque les données sont quantitatives, cas d'un arbre de régression) ou utilisés pour un « vote » à la majorité (pour des données qualitatives, cas d'un arbre de classification). C'est la combinaison de ces multiples estimateurs « indépendants » qui permet de réduire la variance. Toutefois, chaque estimateur est entrainé avec moins de données. En pratique, la méthode de bagging donne d'excellents résultats (notamment sur les arbres de décision utilisés en « forêts aléatoires »).

3.3 Le modèle XGboost

XGBoost signifie eXtreme Gradient Boosting.Il s'agit d'un algorithme de Gradient boosting. Pour Rappel, le Gradient boosting est un algorithme particulier de Boosting. Le Boosting consiste à assembler plusieurs « weak learners » pour en faire un « strong learner », c'est-à-dire assembler plusieurs algorithmes ayant une performance peu élevée pour en créer un beaucoup plus efficace et satisfaisant. L'assemblage de « weak learners » en « strong learner » se fait par l'appel successif de ceux-ci pour estimer une variable d'intérêt. Pour le cas de la classification, il offre la particularité suivante : l'actualisation des poids se calculera de la même façon que la descente de gradient stochastique. La plupart du temps il utilise des algorithmes d'Arbre de Décision (Spécialement ceux utilisant l'algorithme CART) comme des « weak learners ». L'algorithme XGBoost est donc un algorithme ensembliste qui agrège des arbres. À chaque itération, le nouvel arbre apprend de l'erreur commise par l'arbre précédent.
La principale différence entre XGBoost et d'autres implémentations de la méthode du Gradient Boosting réside dans le fait que XGBoost est informatiquement optimisé pour rendre les différents calculs nécessaires à l'application d'un Gradient Boosting rapide. Plus précisément, XGBoost traite les données en plusieurs blocs compressés permettant de les trier beaucoup plus rapidement ainsi que de les traiter en parallèle. L'algorithme XGboost offre des fonctionnalités telles que :

- **Le Calcul distribué ;**
- **La parallélisation ;**
- **Le Calcul out-of-core ;**
- **L'optimisation du cache.**

Mais les avantages de XGBoost ne sont pas uniquement liés à l'implémentation de l'algorithme, et donc à ses performances, mais aussi aux divers paramètres que celui-ci propose. En effet XGBoost propose un panel d'hyperparamètres très important lui offrant l'avantage de :

- **Régularisation :** ces paramètres lui permettent d'éviter le sur-apprentissage ;
- **Flexibilité élevée** ;
- **Gestion des valeurs manquantes :** l'algorithme possède un mécanisme interne de gestion des valeurs manquantes ;
- **Élagage des arbres :** XGBoost effectue des fractionnements jusqu'à la profondeur max_depth spécifiée, puis commence à élaguer l'arbre à l'envers et supprime les fractionnements au-delà desquels il n'y a aucun gain positif ;
- **Validation croisée intégrée :** une validation croisée peut être exécutée à chaque itération du processus de boost et il est donc facile d'obtenir le nombre optimal exact d'itérations de boost en une seule exécution ;
- **Continuité sur le modèle existant :** un modèle XGBoost peut être entrainé à partir de sa dernière itération de l'exécution précédente. Cela peut être un avantage significatif dans certaines applications spécifiques.

Les aspects mathématiques encadrant cet algorithme sont présentés en annexes.

3.4 Le Réseau de neurone

Les réseaux de neurones sont inspirés du fonctionnement des neurones biologiques. Il s'agit d'un système artificiel capable d'apprendre par l'expérience. Le réseau de neurone le plus simple est appelé le perceptron. Ce dernier est composé d'une seule couche et utile pour la compréhension du concept. Néanmoins le perceptron est limité dans la résolution des problèmes et une alternative est l'usage d'un perceptron multicouche. Le perceptron multicouche est organisé en trois partie :

- **La couche d'entrée (input layer) :** un ensemble de neurones qui portent le signal d'entrée. Tous les neurones de cette couche sont ensuite reliés à ceux de la couche suivante ;
- **La couche cachée (hidden layer)** ou plus souvent Les couches cachées (couche

cachée 1, couche cachée 2, ...). Il s'agit du cœur du système. Ces couches implémentent les relations entre les variables ;

- **La couche de sortie (output layer) :** cette couche représente le résultat final du réseau à savoir sa prédiction.

Le mécanisme d'apprentissage repose sur **la descente de gradient**. Cette méthode consiste à trouver les minima d'une fonction continue et différentiable presque partout, en supposant que le gradient de cette dernière est facilement calculable.
Afin de déterminer le résultat final, la couche de sortie utilise une fonction d'activation dont les principales sont : La fonction sigmoïde, le Soft Max, le Rectifier Linear Unit (ReLU) et la tangente hyperbolique.
Dans le cadre de notre travail nous utiliserons l'algorithme MLPClassifier qui implémente un réseau de perceptron multicouche, nous utiliserons la fonction d'activation sigmoïde.

3.5 Les arbres de décision

L'algorithme d'arbre de décision est inspiré de la théorie des graphes, En théorie des graphes, un est un graphe non orienté, acyclique et connexe. L'ensemble des nœuds se divise en trois catégories :

- Nœud racine (l'accès à l'arbre se fait par ce nœud),
- Nœuds internes : les nœuds qui ont des descendants (ou enfants), qui sont à leur tour des nœuds ;
- Nœuds terminaux (ou feuilles) : nœuds qui n'ont pas de descendant.

Il s'agit des méthodes d'apprentissage non paramétriques utilisées pour des problèmes de classification et de régression. Ce modèle prédit les valeurs de la variable cible, en se basant sur un ensemble de séquences de règles de décision déduites à partir des données d'apprentissage. L'arbre approxime donc la cible par une succession de règles if-then-else. Malgré l'inconvénient majeur des arbres de décision à savoir le surapprentissage, ils présentent certains avantages :

- Ils sont simples à comprendre et à visualiser ;
- Ils nécessitent peu de préparation des données (normalisation, etc.) ;
- Ils sont capables de traiter des problèmes multi-classe ;
- Modèle en boîte blanche : le résultat est facile à conceptualiser et à visualiser.

Il existe plusieurs implémentations dont les plus utilisées sont ID3, C4.5, C5 et CART. Ces implémentations diffèrent par le ou les critères de segmentation utilisés, par les mé-

thodes d'élagages implémentées, par leur manière de gérer les données manquantes dans les prédicteurs.
Le principe de construction est le suivant : au départ, les points de la base d'apprentissage sont tous placés dans le nœud racine. Une des variables de description des points est la classe du point ; cette variable est dite « variable cible ». La variable cible peut être catégorielle (problème de classement) ou à valeurs réelles (problème de régression). Chaque nœud est coupé (**opération split**) donnant naissance à plusieurs nœuds descendants. Un élément de la base d'apprentissage situé dans un nœud se retrouvera dans un seul de ses descendants. L'arbre est construit par partition récursive de chaque nœud en fonction de la valeur de l'attribut testé à chaque itération (**top-down induction**). Le critère optimisé est l'homogénéité des descendants par rapport à la variable cible. La variable qui est testée dans un nœud sera celle qui maximise cette homogénéité. Le processus s'arrête quand les éléments d'un nœud ont la même valeur pour la variable cible (homogénéité). Les feuilles de l'arbre spécifient les classes.

3.6 Le modèle Random Forest

Dans le cadre de la classification, les Random Forest utilisent un vote majoritaire pour prédire les classes en fonction de la partition des données provenant de plusieurs arbres de décision. Ces algorithmes de classification réduisent la variance des prévisions d'un seul arbre de décision , améliorant ainsi leurs performances. Ils combinent de nombreux arbres de décisions dans une approche de type bagging. Ils effectuent un apprentissage en parallèle sur de multiples arbres de décision construits aléatoirement et entraînés sur des sous-ensembles de données différents. Le nombre idéal d'arbres, qui peut aller jusqu'à plusieurs centaines voire plus, est un paramètre important : il est très variable et dépend du problème. Concrètement, chaque arbre de la forêt aléatoire est entrainé sur un sous ensemble aléatoire de données selon le principe du bagging, avec un sous ensemble aléatoire de variables selon le principe des « projections aléatoires ». Ils ont la particularité d'être un des classifieurs les plus efficaces « out-of-the-box » c'est-à-dire nécessitant peu de prétraitement des données.
Les avantages qui motivent l'utilisation de ces algorithmes sont les suivantes :

- ✧ • La formation et la prédiction sont très rapides, en raison de la simplicité des arbres de décision sous-jacents. De plus, les deux tâches peuvent être directement parallélisées, car les arborescences individuelles sont des entités entièrement indépendantes ;
- ✧ • Les Random Forest permettent une classification probabiliste : un vote majoritaire parmi les estimateurs donne une estimation de la probabilité (accessible dans Scikit-Learn avec la méthode predict_proba ()) ;

- • Le modèle non paramétrique est extrêmement flexible et peut donc bien fonctionner sur des tâches sous-ajustées par d'autres estimateurs.

Dans le cadre de notre travail, nous avons utilisé un ensemble optimisé d'arbres de décision aléatoires implémenté dans l'estimateur RandomForestClassifier de scikit learn et qui prend en charge automatiquement toute la randomisation.

3.7 Le modèle KNN

L'algorithme de K Nearest Neighbors (KNN) est une méthode de classification supervisée, utilisée aussi bien pour la régression que pour la classification. Pour effectuer une prédiction, cet algorithme a besoin :

- • Un ensemble de données d'apprentissage D ;
- • Une fonction de distance d ;
- • Et un entier k.

Pour tout nouveau point de test x, pour lequel il doit prendre une décision, l'algorithme recherche dans D les k points les plus proches de x au sens de la distance d, et attribue x à la classe qui est la plus fréquente parmi ces k voisins. KNN n'a pas besoin de construire un modèle prédictif. Ainsi, pour KNN il n'existe pas de phase d'apprentissage proprement dite. C'est pour cela qu'on le catégorise parfois dans le Lazy Learning. Pour effectuer la prédiction, si KNN est utilisé pour la régression, c'est la moyenne (ou la médiane) des variables des plus proches observations qui servira pour la prédiction. Par contre dans le cadre de la classification, c'est le mode des variables des plus proches observations qui servira pour la prédiction.
Le choix du nombre de voisin optimal varie en fonction du jeu de données. En règle générale, moins on utilisera de voisins (un nombre k petit) plus on sera sujette au sous apprentissage (underfitting).Par ailleurs, plus on utilise de voisins (un nombre K grand) plus, sera fiable notre prédiction. Toutefois, si on utilise k nombre de voisins avec k=n et n étant le nombre d'observations, on risque d'avoir du overfitting et par conséquent un modèle qui se généralise mal sur des observations qu'il n'a pas encore vues. L'inconvénient majeur de cet algorithme est le fait qu'il est couteux en mémoire et lent pour un jeu de données de taille élevée. En effet,il doit garder en mémoire l'ensemble des observations pour pouvoir effectuer sa prédiction. Néanmoins ils sont d'une grande simplicité de compréhension.

3.8 Le modèle AdaBoost

Adaboost pour adaptive boosting est un algorithme de boosting. C'est un principe qui regroupe de nombreux algorithmes qui s'appuient sur des ensembles de classifieurs binaires : le boosting optimise leurs performances. Le principe est issu de la combinaison de classifieurs (appelés également hypothèses). Par itérations successives, la connaissance d'un classifieur faible - weak classifier - est ajoutée au classifieur final - strong classifier. La spécificité de Adaboost réside dans le fait qu'à chaque étape, le calcul de la nouvelle pondération est mené de manière à ce que le nouvel ensemble d'apprentissage soit mal classé par la combinaison linéaire des classificateurs précédents. Adaboost est donc un algorithme d'optimisation. Son principe peut se résumer comme suit :

- • Associer une distribution de poids à tous les exemples de la base d'apprentissage ;
- • Cette distribution change après chaque itération ;
- • Des poids plus importants sont affectés aux exemples qui sont mal classifies par l'itération précédente.

3.9 Le modèle SVM

Le « support vector machines » fait l'objet d'un grand intérêt théorique et de nombreuses publications. Il connaît un engouement comparable à celui connu auparavant, des réseaux de neurone, avec lesquels ils sont parfois comparés. Les fondements théoriques des SVM sont toutefois plus proches de ceux des méthodes statistiques classiques, s'agissant en effet, d'une sorte de généralisation de l'analyse discriminante linéaire. On distingue deux cas de SVM, selon la configuration des donnés.
Le premier aui est assez théorique et est basé sur les travaux de Vladimir Vapnik à partir de 1995, est un cas assez théorique et stipule un classement linéairement séparé des observations.
Le second cas, appelé souvent "cas non-séparé", est plus pratique. Cependant, sa compréhension dans les détails nécessite un développement mathématique, nottament sur les notions d'hyperplan, et ne saurait pouvoir être contenu dans ce document qui se veut synthétique. Nous renvoyons au lecteur au livre de Stéphane TUFFERY à la page 622 pour plus de détails.

Chapitre

4

Etude comparative des techniques

Enfin, ce chapitre entre dans le vif du sujet et à pour objectif de répondre avec la pratique à la problématique posée. Il sera alors question de présenter les critères retenus pour trancher entre lequel des différents modèles présentés au chapitre précédent sera retenu. A la suite de cette étape, nous passerons aux méthodes d'optimisation des modèles retenus avant de présenter les résultats obtenus. Enfin, ce chapitre se focalisera sur le modèle retenu au regard des critères énumérés.

4.1 Critères de choix des modèles 25
4.2 Optimisation des modèles 26
4.3 Résultats 26
4.4 Modèle retenu : XGBoost 29
.1 Statistiques descriptives iv

4.1 Critères de choix des modèles

Plusieurs critères peuvent être utilisés dans le cadre d'une classification binaire pour comparer les modèles. Le modèle idéal dans notre étude sera d'abord celui qui est capable de se généraliser sur de nouvelles observations. Nous éviterons donc les modèles qui overfit. Parmi les modèles qui se généralisent bien, nous serons plus attentifs à celui qui maximise la sensibilité. En effet, stratégiquement, notre modèle n'a pas intérêt à prédire qu'un client va continuer à utiliser le service alors que celui-ci va abandonner. Cela est plus dommageable pour la banque qui a plutôt intérêt à bien repérer les clients susceptibles d'abandonner afin de les cibler par une politique adéquate. Pour rappel, la sensibilité désigne taux de vrais positifs, c'est à dire la proportion de départ que l'on a correctement identifiés. C'est la capacité de notre modèle à bien détecter l'attrition d'un client. Autres ces deux critères primordiaux nous observerons la courbe de ROC, l'AIC etc.

4.2 Optimisation des modèles

Après avoir implémenté les modèles avec l'ensembles des variables, nous sélectionnerons les variables les plus pertinentes pour expliquer l'attrition client. Cette étape, qui est l'étape de la feature selection, offre l'avantage de :

- Réduire l'overfitting : moins de données redondantes signifie moins de possibilité de prendre des décisions basées sur des données / bruits redondants ;
- Améliore la précision : Moins de données trompeuses signifie que la précision de la modélisation s'améliore ;
- Réduit le temps d'entraînement : moins de données signifie que les algorithmes s'entraînent plus rapidement et que la Banque aura besoin de peut d'informations pour prédire l'attrition..

Pour ce faire, nous utiliserons la classe selectKbest du package scikitlearn. Cette classe effectue un test statistique univarié entre chaque variable et la variable d'interêt (test du chi2, ou ANOVA dans certains cas) et affecte un score pour chaque variable. Finalement on ne retient que les k variables ayant les scores les plus élevés dans notre modèle.

Aussi, nous utiliserons la validation croisée pour présenter toutes les données aux modèles afin d'éviter l'overfitting.
• Transformation : normalisation
• Gridsearch

4.3 Résultats

Aprés compilation des modèles retenus, nous obtenons les résultats suivants avec les différents indicateurs de comparaison. D'une manière générale, nous pouvons dire que les modèles donnent des taux acceptables au niveau de la sensibilité, de la précision et du F1-score. Cependant un autre critère important dans la modélisation (le sur-apprentissage) nous a permis de réduire la sélection à quatre (4).
Au total, nous avons testé neuf (9) modèles et avons regroupons les sorties selon trois catégories :

Tableau 4.1 – *Résultats sur les réseaux de neurones*

Résultats sur les réseaux de neurones				
	precision	**recall**	**f1-score**	**support**
0	0,78	0,63	0,69	187
1	0,92	0,96	0,94	819
accuracy			0,90	1006
macro avg	0,85	0,79	0,82	1006
weighted avg	0,89	0,90	0,89	1006

Source : Base de données, nos calculs.

Tableau 4.2 – *Résultats sur les modèles ensemblistes*

Résultats sur les modèles ensemblistes				
	precision	**recall**	**f1-score**	**support**
Bagging				
0	0,73	0,63	0,67	187
1	0,92	0,95	0,93	819
accuracy			0,89	1006
macro avg	0,82	0,79	0,80	1006
weighted avg	0,88	0,89	0,88	1006
XGBOOST (modèle retenu)				
0	**0,83**	**0,64**	**0,72**	**187**
1	**0,92**	**0,97**	**0,95**	**819**
accuracy			**0,91**	**1006**
macro avg	**0,88**	**0,80**	**0,83**	**1006**
weighted avg	**0,90**	**0,91**	**0,90**	**1006**
Randomized Decision Tree				
0	0,880	0,390	0,541	187
1	0,876	0,988	0,929	819
accuracy			0,877	1006
macro avg	0,878	0,689	0,735	1006
weighted avg	0,877	0,877	0,857	1006
Addaboost				
0	0,794	0,535	0,639	187
1	0,901	0,968	0,933	819
accuracy			0,888	1006
macro avg	0,847	0,752	0,786	1006
weighted avg	0,881	0,888	0,879	1006

Source : Base de données, nos calculs.

Tableau 4.3 – *Résultats sur les modèles standards*

Résultats sur les modèles standards				
Indicateurs	**precision**	**recall**	**f1-score**	**support**
Régression Logistique				
0	0,77	0,46	0,58	187
1	0,89	0,97	0,93	819
accuracy			0,87	1006
macro avg	0,83	0,71	0,75	1006
weighted avg	0,87	0,87	0,86	1006
Arbre de décision				
0	0,66	0,60	0,63	187
1	0,91	0,93	0,92	819
accuracy			0,87	1006
macro avg	0,79	0,77	0,78	1006
weighted avg	0,86	0,87	0,87	1006
KNN				
0	0,77	0,56	0,65	187
1	0,91	0,96	0,93	819
accuracy			0,89	1006
macro avg	0,84	0,76	0,79	1006
weighted avg	0,88	0,89	0,88	1006
SVM				
0	0,83	0,53	0,65	187
1	0,90	0,98	0,94	819
accuracy			0,89	1006
macro avg	0,87	0,76	0,79	1006
weighted avg	0,89	0,89	0,88	1006

Source : Base de données, nos calculs.

4.4 Modèle retenu : XGBoost

Après avoir normaliser et sélectionner les cinq variables pertinentes pour modéliser l'attrition, quatre modèles sont éligibles car ils n'overfittent pas ou plus : Il s'agit de **la régression logistique, du support vecteur machine, du Random Decision Tree et du XGboost**. Ces modèles ont les caractéristiques suivantes :

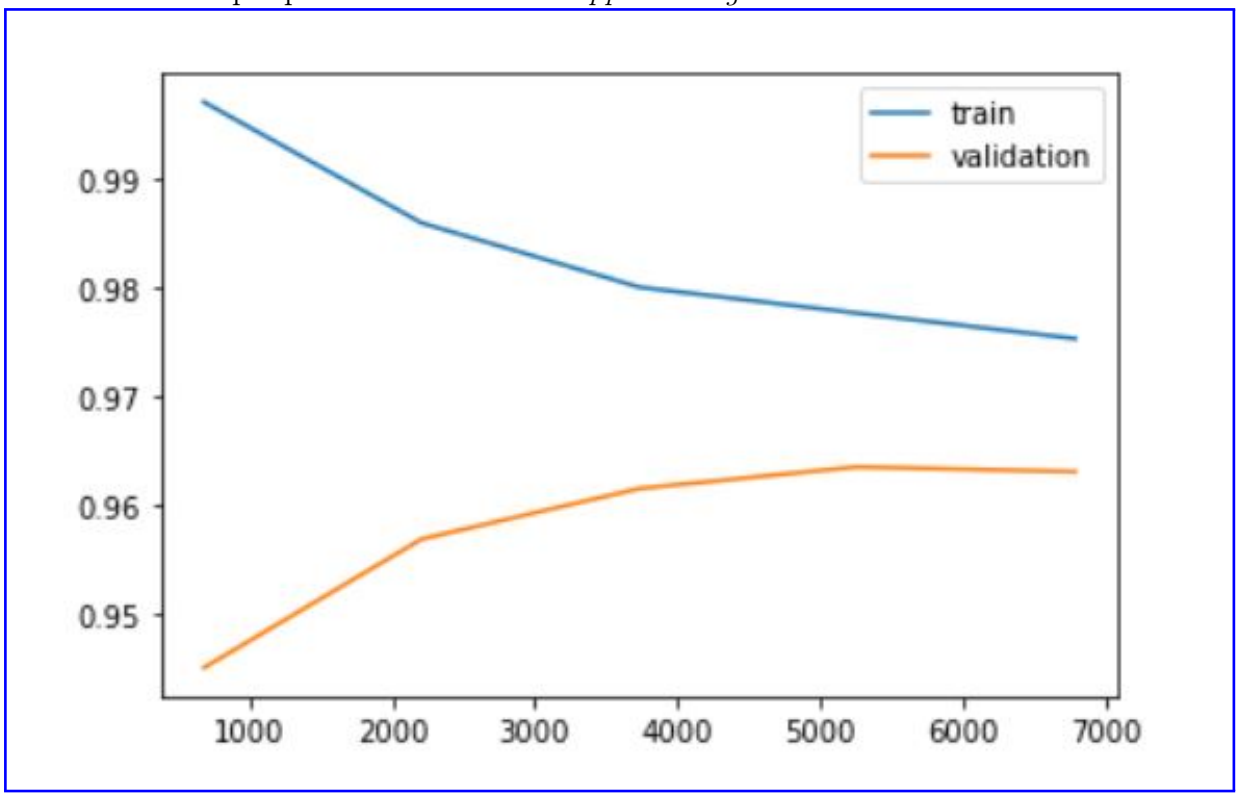
Tableau 4.4 – *Les résultats sur les modèles sans surapprentissage*

Modèles sans surapprentissage				
Critères	Logistique	SVM	XGBoost	Random Forest
Accuracy	0,87	0,89	0,91	0,88
Sensibilité	0,89	0,90	0,92	0.88
Area Under Curve (AUC)	0,71	0,75	0,80	0,69

Source : Base de données, nos calculs.

Le modèle XGboost étant celui qui maximise les trois critères, il est retenu et ces paramètres seront optimisés pour améliorer ces performances. Le graphe suivant montre que le modèle ne fait pas du sur-apprentissage et tant à un niveau stable de performance aussi bien dans la base test que dans la base d'apprentissage au fur et à mesure que les tailles des échantillons augmentent. En plus de ce graphe, on peut voir la courbe de ROC

Graphique 4.1 – *Courbe d'apprentissage du modèle XGBoost*

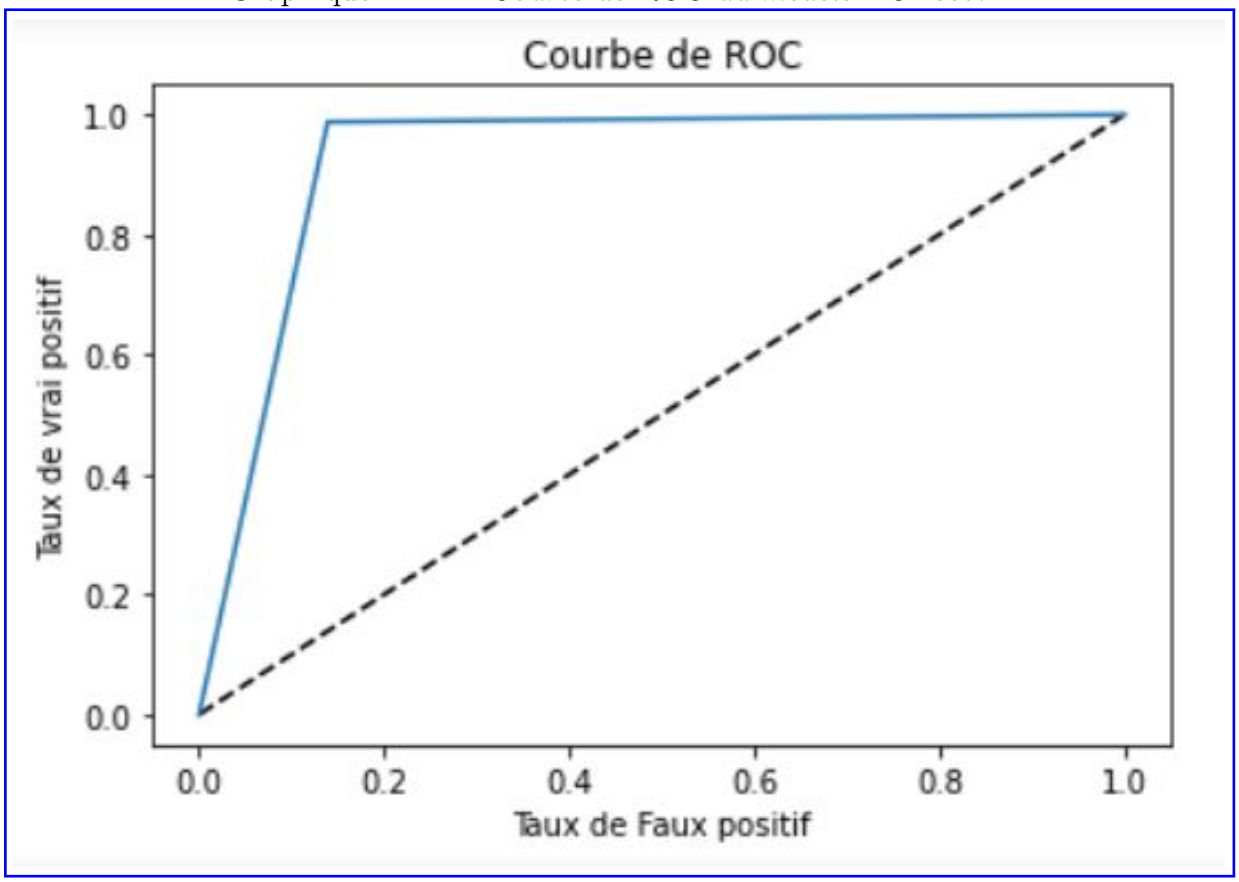


Source : Base de données, nos calculs.

montrant la capacité prédictive du modèle et donc la capacité du modèle à prédire l'attrition d'un client.

Le pouvoir prédictif du modèle est de 92,40%.

Graphique 4.2 – *Courbe de ROC du modèle XGBoost*



Source : Base de données, nos calculs.

POSTAMBULE

Difficultés et Limites de l'étude

Au cours de la réalisation de ce projet, certes il a été d'une grande utilité dans la mesure où il nous a permis de maitriser en grande partie les techniques du DataMining, nous nous sommes confrontés à quelques difficultés.

- ☞ Au cours du travail, nous avons rencontré des difficultés à la fois d'ordre théorique et pratique. Au plan théorique, la compréhension des théories mathématiques sous-jacentes aux modèles ainsi que les techniques de réduction du nombre de variables dans le modèle se sont avérées chrono-phages.
- ☞ Pratiquement, le paramétrage manuel conformément aux théories des modèles et de l'incompatibilité de certains paramètres pour chaque modèle surtout pour l'algorithme XGboost fut fastidieux, cherchant une alternative nous avons voulu optimiser les paramètres du modèle retenu à l'aide de Gridsearch. Cependant le temps d'exécution est excessivement long.
- ☞ En outre, dans la recherche d'un algorithme de discrétisation dédiée à la prédiction, nous sommes tombés sur l'algorithme MDLPC de Fayyad & Irani, qui dans la mise en œuvre s'est révélé très compliqué à implémenter. Bien qu'il était très indiqué pour notre type d'étude, nous nous sommes rabattu à des techniques standards.

S'agissant des limites de l'étude, la première chose à remarquer est le nombre non consistant de nos données. Cette limite, si elle était résolue, bien que le modèle retenu ne fait pas ou trop de sur-apprentissage avec un pouvoir prédictif élevé, pourrait conduire à des taux de sensibilité, de précision et à un F1-score plus élevés. Cela se note aussi dans la courbe d'apprentissage du modèle XGBoost, en particulier.

Conclusion

Bilan de l'étude

En définitive, le travail a consisté de manière globale à mettre en œuvre un modèle de prédiction de l'attrition des clients d'une banque de la place. Pour atteindre cet objectif, nous avons fait appel aux connaissances théoriques et pratiques acquises au cours de DataMining sous la supervision et l'accompagnement du professeur. Les techniques exploitées, nous ont permis de sélectionner, d'abord, les modèles qui ne sur-apprennent pas à l'aide des courbes de sur-apprentissage et, ensuite, de retenir parmi ceux-ci le modèle qui répond au mieux à nos critères.

Perspectives

En termes de perspectives de l'étude, nous proposons une plateforme de décision, qui consisterait en mettre en place un système de reporting permettant à un gestionnaire de la clientèle de saisir les information d'un client ou d'un groupe de clients (leurs valeurs sur les cinq variables retenues) afin de voir si il ya menace d'attrition ou non et pouvoir cibler ces clients pour mener des politiques de rétention. Ces politiques peuvent être d'ordre de publicité, de proposition de nouveaux services, améliorer le marketing, faire baisser les taux d'intérêt ou le coût de certains services pour certains clients. La soumission des données clients peut aussi être automatique, sans nécessité un gestionnaire, à partir d'un DataWareHouse et qui lance une alerte lorsque le modèle prédit qu'un client risque l'attrition. Pour l'outil de reporting, des outils comme **R-Shiny** ou **Flask** (Python) peuvent être utilisé.

Bilan sur le cours

De manière générale, le cours de DataMining a été très intéressant dans son ensemble. Il nous a été bénéfique sur tous les plans et nous a permis de comprendre les concepts essentiels et actuels du Machine Learning. En outre, il nous a permis de comprendre et de maitriser le paradigme de Python qui est aujourd'hui essentiel dans la carrière d'un statisticien. En plus, bien que le professeur ait eu à nous apprendre plusieurs concepts directement en cours, il l'a aussi fait indirectement à travers ce projet. Car ce projet nous a permis de maitriser d'autres modèles et de comprendre d'autres algorithmes à travers les recherches que nous avons eu à faire.

Références Bibliographiques

Ouvrages

[1] Abdou Aziz NDIAYE, 2020-2021, *Cours de DataMining - ENSAE-Dakar*, Cours.

[2] Andrea Pietracaprina et al., 2015, *Machine learning techniques for customer churn prediction in banking environments* , Cours.

[3] PARTEEK Bhatia, 2019, *Data Mining and Data Warehousing Principles and Practical Techniques*, Livre.

[4] Stéphane TUFFERY, 2012, *Data Mining et Statistique décisionnelle, L'intélligence des donnée*, Livre, quatrième édition.

Sites web

[1] 2021,*https ://www.datacorner.fr/xgboost/*

[2] 2021,*https ://datafuture.fr/post/faire-tourner-xgboost-sous-r/*

[3] 2021,*https ://ichi.pro/fr/comment-fonctionne-xgboost-128143693994154*

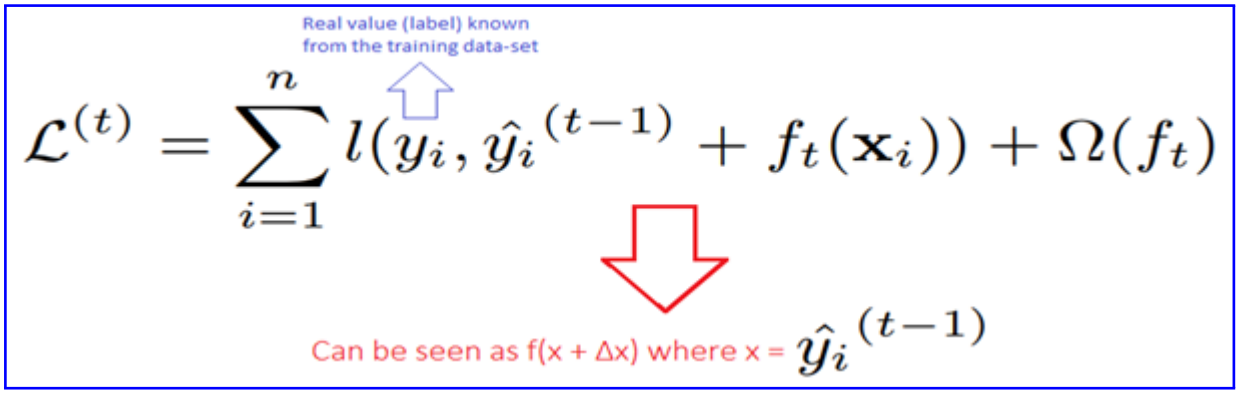
[4] 2021, *https ://www.kdnuggets.com/2019/05/churn-prediction-machine-learning.html*

[5] 2021, *https ://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/*

Annexes

Le modèle XGBoost et sa spécification mathématique

Mathématiquement, la fonction objectif (fonction de perte et régularisation) à l'itération t à minimiser est la suivante :



Afin d'utiliser les techniques d'optimisation traditionnelles, nous effectuons une transformation de la fonction objective d'origine en une fonction du domaine euclidien. Une approximation de Taylor du second ordre donne :

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2$$

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n}[l(y_i, \hat{y}^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2}h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t)$$

Avec ,

$$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}) \text{ and } h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$$

En soustrayant la constante de part et d'autre, nous obtenons une forme simplifiée :

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} [g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t)$$

La fonction ci-dessus est une somme de fonctions quadratiques simples d'une variable et peut être minimisé en utilisant des techniques d'optimisation classiques. En pratique, l'algorithme fonctionne de la manière suivante :

• Débute avec une racine unique ;

• Itérer pour toute les variables et toutes les valeurs par variables ;

• A chaque itération, calculer le gain possible : Gain = perte (instances pères) - (perte (branche gauche) + perte (branche droite)) Le gain pour le meilleur partage doit être positif (et > au paramètre min_split_gain), sinon le développement de la branche est arrêté.

Pour le cas de la classification binaire, le logarithme de la fonction objective est la suivante :

$$y \ln(p) + (1 - y) \ln(1 - p) \text{ where } p = \frac{1}{(1 + e^{-x})}$$

LE TEST DE GRUBBS

Test de Grubbs:

Principe: Le test de Grubbs permet de détecter les valeurs aberrantes en termes de dispersion de moyennes. Le principe de ce test est de comparer les valeurs absolues des écarts réduits, en d'autres termes :

$$G_i = \frac{\max|X_i - \bar{X}|}{S_X}$$

x_i : point de mesure le plus éloignée de la moyenne

x_{bar}: moyenne des mesures

S_X : écart type des mesures

Si la valeur G_i est supérieure à la valeur critique G_c donnée dans la table ci-dessous, alors la mesure du laboratoire incriminé est supprimée. Ce test ne permet de tester qu'un seul point en même temps, par conséquent, on répète ce test tant que G_i est supérieur à la valeur critique.

.1 Statistiques descriptives

Tableau 5 – *L'attrition selon le sexe des clients*

L'attrition selon le sexe des clients		
Attrition_Flag/Sexe	**Féminin**	**Masculin**
Attrited Customer	929	695
Existing Customer	4405	4031

Source : Base de données, nos calculs.

Tableau 6 – *L'attrition selon la situation matrimoniale*

L'attrition selon la situation matrimoniale				
Attrition_Flag	**Divorced**	**Married**	**Single**	**Unknown**
Attrited Customer	121	706	668	129
Existing Customer	624	3936	3261	615

Source : Base de données, nos calculs.

Tableau 7 – *L'attrition selon le type de la carte*

L'attrition selon le type de la carte				
Attrition_Flag	**Blue**	**Gold**	**Platinum**	**Silver**
Attrited Customer	1516	21	5	82
Existing Customer	7857	95	15	469

Source : Base de données, nos calculs.

Tableau 8 – *L'attrition selon la catégorie de revenu*

L'attrition selon le revenu		
Revenu/Attrition_Flag	**Attrited Customer**	**Existing Customer**
$120K +	126	598
$40K - $60K	270	1510
$60K - $80K	188	1200
$80K - $120K	242	1275
Less than $40K	611	2936
Unknown	187	917

Source : Base de données, nos calculs.

Tableau 9 – *L'attrition selon le niveau d'éducation*

L'attrition selon le niveau d'éducation		
Attrition/Niveau etude	**Attrited Customer**	**Existing Customer**
College	154	857
Doctorate	95	351
Graduate	487	2624
High School	305	1692
Post-Graduate	92	421
Uneducated	235	1237
Unknown	256	1254

Source : Base de données, nos calculs.

Tableau 10 – *Les variables quantitatives et leurs satatistiques*

Les variables quantitatives et leurs satatistiques						
Variables	**mean**	**std**	**min**	**50%**	**75%**	**max**
Customer_Age	45,00	5,05	40	45	49	51
Dependent_count	3,60	0,89	3	3	4	5
Months_on_book	34,80	8,58	21	36	39	44
Total_Relationship_Count	4,60	1,14	3	5	5	6
Months_Inactive_12_mon	1,60	1,34	1	1	1	4
Contacts_Count_12_mon	1,20	1,30	0	1	2	3
Credit_Limit	6478,80	4007,75	3313	4716	8256	12691
Total_Revolving_Bal	831,60	1028,07	0	777	864	2517
Avg_Open_To_Buy	5647,20	4233,43	796	4716	7392	11914
Total_Amt_Chng_Q4_Q1	1,81	0,55	1,335	1,541	2,175	2,594
Total_Trans_Amt	1261,80	391,34	816	1171	1291	1887
Total_Trans_Ct	28,60	9,32	20	28	33	42
Total_Ct_Chng_Q4_Q1	2,50	0,76	1,625	2,333	2,5	3,714
Avg_Utilization_Ratio	0,19	0,32	0	0,061	0,105	0,76

Source : Base de données, nos calculs.

Table des matières

Décharge i

Sommaire iii

Liste des graphiques iv

Liste des tableaux v

Résumé vi

Preambule 2

Contextualisation 2

Introduction 3

I Etude préliminaire : Exploration et préparation des données 5

1 Présentation et traitement de la base de données 6

1.1 Description de la base de données 6

1.1.1 Variables socio-démographiques (6) 6

1.1.2 Variables bancaires (13) 7

1.2 Traitement préliminaire des données 8

1.2.1 Valeurs manquantes 9

1.2.2 Doublons 9

1.2.3 Valeurs extrêmes et valeurs aberrantes 9

1.2.4 Discrétisation 10

1.2.5 Catégorisation 10

2 Profilage des clients et choix des variables 11

2.1 Analyse univariée 11

2.1.1 Attrition des clients 11

2.1.2 Caractéristiques socio-professionnelles des clients 12

2.1.3 Relation des clients avec la banque 12

2.2 Analyse multivariée 13
2.3 Détection des variables les plus discriminantes 13
2.3.1 Présentation des critères de choix 14
2.3.2 Présentation des variables retenues 15

II Modélisation et présentation des résultats 17

3 Les différentes techniques de modélisation 18
3.1 Le modèle logistique 18
3.2 Le modèle Bagging 19
3.3 Le modèle XGboost 19
3.4 Le Réseau de neurone 20
3.5 Les arbres de décision 21
3.6 Le modèle Random Forest 22
3.7 Le modèle KNN 23
3.8 Le modèle AdaBoost 24
3.9 Le modèle SVM 24

4 Etude comparative des techniques 25
4.1 Critères de choix des modèles 25
4.2 Optimisation des modèles 26
4.3 Résultats 26
4.4 Modèle retenu : XGBoost 29

Postambule 33

Difficultés et Limites de l'étude 33

Conclusion 34

Références Bibliographiques i

Annexes ii
.1 Statistiques descriptives iv

Table des matières viii