

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Saad Dahleb de Blida 1



Faculté des sciences

Département d'Informatique

Domaine : Mathématique et informatique

Filière : Informatique

Spécialité : Ingénierie de Logiciel

Mémoire présenté pour obtenir diplôme de master 2

Soutenu Le 7 juillet 2022

Conception et réalisation d'un system de prédiction de client churn dans le secteur de télécommunications

Réalisé par

— Hocine Abir

Devant le jury composé :

— **Président** : Mr.Hammouda MAA à l'Université de Blida 1

— **Examineur** : Mm.Djedar MCB à l'Université de Blida 1

— **Encadrant** : Mm.Daoud Hayat MAB à l'Université de Blida 1

Remerciements

Notre louange va surtout à Dieu Tout-Puissant, en qui j'ai foi et qui a récompensé moi avec ce travail présent.

Je tiens à exprimer ma profonde gratitude à mon directeur de thèse, le Dr Daoud Hayat, pour l'aide qu'elle m'a apportée dans la réalisation de ce document de recherche.

Je tiens à remercier affectueusement : Fatma Salmi ma mère, et tous les membres de la famille, les amis et tous ceux qui m'ont soutenu, de près ou de loin, tout au long des années d'études. Je vous remercie sincèrement, parce que vous avez toujours été présents. Que cette œuvre soit un témoignage de ma gratitude et de mon profond respect.

Dédicaces

À ma mère aimante ;

Mes chers grands-parents, mes tantes, et mes oncles qui ont ma gratitude éternelle, que Dieu vous donne longue vie, santé et bonheur.

A mes chères sœurs Aicha, Amina, Chaimaa, Maroua, Rihame, Kenza mes frères Abdelouaheb, Allaedinne, Abdennour qui sans eux je ne serais pas où je suis aujourd'hui.

Aussi, à mon beau-frère Madjid et à leurs beaux enfants Tasnime, Abdelmalek, Djihane.

À tous mes amis et à tous ceux qui m'ont marqué au cours de ce glorieux voyage.

Résumé

Les entreprises de télécommunications dans le monde entier s'intéressent au lien existant entre la satisfaction du client et les revenus de l'entreprise. Le client churn dans le domaine des télécommunications fait référence à un client qui cesse sa relation avec l'entreprise et il est également probable que ce client rejoindra une entreprise concurrente.

En effet, ces entreprises s'appuient sur deux stratégies principales pour générer plus chiffre d'affaires :

Revaloriser les clients existants et augmenter la fidélisation du client.

Ce rapport a été élaboré dans le cadre du projet de fin d'études pour l'obtention du diplôme d'ingénieur de logiciel, ce projet a comme objectif d'identifier les clients churn qui sont plus proches d'abandonner leur opérateur de télécommunications en utilisant des techniques de la machine learning pour la prédiction du churn.

Mots clés : Prédiction du churn, Télécommunication, Désabonnement des clients

ملخص

تهتم شركات الاتصالات حول العالم بالصلة بين رضا العملاء وإيرادات الشركة. يشير العميل المضطرب في مجال الاتصالات إلى العميل الذي يتوقف عن علاقتة بالشركة ومن المحتمل أيضًا أن ينضم هذا العميل إلى شركة منافسة.

في الواقع، تعتمد هذه الشركات على استراتيجيتين رئيسيتين لتوليد المزيد من الإيرادات :
ترقية العملاء الحاليين وزيادة ولاء العملاء.

تم إعداد هذا التقرير كجزء من مشروع تخرج هندسة البرمجيات، ويهدف هذا المشروع إلى تحديد العملاء المضطربين الذين يقربون من التخلي عن مشغل الاتصالات الخاص بهم باستخدام تقنيات التعلم الآلي.

الكلمات المفتاحية : التنبؤ، الاتصالات بإلغاء اشتراك العميل

Abstract

Telecom companies around the world are interested in the link between customer satisfaction and company revenue. The churn customer in the field of telecommunications refers to a customer who ceases his relationship with the company and it is also likely that this customer will join a competing company.

Indeed, these companies rely on two main strategies to generate more revenue : Upgrade existing customers and increase customer loyalty.

This report was prepared as part of the Software Engineering Graduation Project, this project aims to identify churn customers who are closer to abandoning their telecom operator by using machine learning techniques for churn prediction.

Keywords :Churn prediction, Telecommunication, churn clients.

Table des matières

1	Introduction Général	14
1.1	Contexte	14
1.2	La problématique	15
1.3	Les objectives	15
1.4	La structure de mémoire	15
2	Clients Churn	17
2.1	Introduction	17
2.2	Gestion de la Relation Client CRM	18
2.2.1	Définition et Principes de CRM	18
2.2.2	Les Trois Parties du CRM	19
2.2.3	Les Objectifs de CRM	20
2.2.4	Définition de E_CRM	22
2.2.5	La Différence entre CRM et E_CRM	23
2.3	Satisfaction de la clientèle	24
2.3.1	Le Client : la personne la plus convoitée	24
2.3.2	Les Cinq pouvoirs du client	25
2.3.3	Satisfaction de la clientèle	25
2.3.4	Les 19 Impacts de la satisfaction client dans la rentabilité des entreprises	26
2.4	Le Churn du Client	28
2.4.1	Valeur à vie du client (Customer Lifetime Value CLV)	28
2.4.2	Définition de Churn	28
2.4.3	Types de Churn	29

2.4.4	Les Déterminants du Churn	31
2.4.5	Le Churn est Coûteux	33
2.4.6	Le Churn est difficile à Gérer	34
2.4.7	Churn vous aide à apprendre à propos de l'entreprise	35
2.5	Conclusion	36
3	L'Apprentissage Automatique (Machine Learning)	37
3.1	Introduction	37
3.2	L'Apprentissage Automatique	37
3.2.1	Définition d'Apprentissage Automatique ML	38
3.2.2	L'Histoire d'Apprentissage Automatique ML	39
3.2.3	Les Types d'Apprentissage Automatique	40
3.2.4	La Nécessité de Machine Learning (ML) dans la Détection de Churn	42
3.2.5	Fonctionnement d'un Système ML pour la Détection de Churn . . .	44
3.3	L'Apprentissage Supervisé	45
3.3.1	La Prédiction	45
3.3.2	Régression ou Classification	46
3.3.3	Les Méthodes d'Apprentissage Supervisé	47
3.4	L'Apprentissage Non Supervisé	49
3.4.1	Clustering	49
3.4.2	Les Méthodes d'Apprentissage Non Supervisé	50
3.4.3	Choix du Nombre de Cluster	54
3.5	Prétraitement des Données	55
3.5.1	Problèmes de Données	55
3.5.2	Techniques de Prétraitement des Données	56
3.6	Méthodes d'équilibrage des données	58
3.6.1	Méthodes de Over_Sampling	59
3.6.2	Méthodes de Under_Sampling	59
3.7	Évaluation des modèles d'apprentissage automatique	60
3.7.1	Train_Test Split	60
3.8	Méthodes d'ensemble	61
3.8.1	La méthode de stimulation (Boosting)	62
3.8.2	Classificateur de vote (Voting Classifier)	63

3.8.3	La methode d'ensachage (Bagging)	64
3.9	Les Métriques d'évaluation	65
3.9.1	Matrice de confusion	65
3.9.2	Précision	66
3.9.3	Rappel	66
3.9.4	F_mesure	66
3.9.5	Taux de succès (exactitude) et taux d'erreur	67
3.9.6	La courbe des caractéristiques d'exploitation du récepteur (Receiver Operating Characteristic Curve ROC)	67
3.10	Conclusion	68
4	Conception	70
4.1	Introduction	70
4.2	Architecture du système	70
4.2.1	Collection des Données	71
4.2.2	Analyse des Données	71
4.2.3	Le Prétraitement des données	79
4.3	Conclusion	82
5	Réalisation et Tests	83
5.1	Introduction	83
5.2	Évaluation des modèles	83
5.2.1	Matrice de confusion	83
5.2.2	Courbe ROC	86
5.3	Discussion des résultats	88
5.4	L'Interface d'application	90
5.4.1	La Page d'accueil	90
5.4.2	La Page des résultats de prédiction	91
5.5	Outils et technologies utilisées	91
5.5.1	Environnement de travail	91
5.5.2	Analyse exploratoire des données	92
5.5.3	Prédiction	93
5.5.4	L'Interface graphique	93

5.6	Conclusion	93
6	Conclusion Général	94
6.1	Conclusion	94
6.2	Perspectives	95

Table des figures

2.1	Le modèle de système de CRM qui comprend tous les types	19
2.2	La différence entre CRM et E_CRM	23
2.3	Les 19 impacts de la satisfaction client	26
2.4	Involontaires vs Volontaires Churn	29
2.5	La taxonomie de churn	30
3.1	Les Types d'Apprentissage Automatique	41
3.2	La structure de base du fonctionnement des algorithmes de détection de churn par l'apprentissage automatique	44
3.3	Le modèle de régression logistique.	47
3.4	Le schéma de K_means	53
3.5	Méthodes d'équilibrage des données	59
3.6	Les opérations de Train_Test Split	61
3.7	Schéma de la méthode de stimulation (Boosting)	62
3.8	Schéma de Classificateur de vote (Voting)	64
3.9	Schéma de la methode d'ensachage (Bagging)	65
3.10	Matrice de Confusion	66
3.11	La courbe des caractéristiques d'exploitation du récepteur	68
4.1	L'organigramme d'architecture du système	71
4.2	Structure de Données	74
4.3	Description des Données	75
4.4	La Distribution de la variable churn	76
4.5	Répartition du churn selon l'ancienneté des abonnées	77
4.6	Matrice des corrélations entre les fonctionnalités	78

4.7	Les corrélations avec la variable cible	78
4.8	Le diagramme d'inertie	80
4.9	Les types des clients	81
4.10	l'Équilibrage des données	82
5.1	Matrice de confusion	84
5.2	Matrice de confusion de la méthode de stimulation	85
5.3	Matrice de confusion de méthode d'ensachage	86
5.4	Courbe ROC	87
5.5	Courbe ROC de la méthode de stimulation	88
5.6	Courbe ROC de la méthode d'ensachage	88
5.7	La Page d'accueil	90
5.8	La Page des résultats de prédiction	91

Liste des tableaux

4.1	Les variables du jeu de données.	73
5.1	Comparaison des modèles	84
5.2	Comparaison des résultats de méthode de stimulation	85
5.3	Comparaison des résultats de méthode d'ensachage	86

Liste des abréviations

CRM Customer Relationship Management

E_CRM Electronique Customer Relationship Management (Electronique de Gestion de la Relation Client)

GRC Gestion de la Relation Client

CLV Customer Lifetime Value (Valeur à vie du client)

ML Machine Learning (L'Apprentissage Automatique)

IT Technologie de l'Information (Information Technology)

RL Régression logistique

RF Random Forest (La Forêt aléatoire)

MML Minimum Message Length

MDL Minimum Description Length

BIC Bayes Information Criterion

AIC Akaike Information Criterion

ROC Receiver Operating Characteristic

AUC l'Area Under the Curve (Surface sous la courbe)

CSV Comma Separated Values

DTC Decision Tree Classifier (Classificateur de l'arbre décisionnel)

SVC Classificateur de difficile vote (Soft Voting Classifier)

HVC Classificateur de doux vote (Hard Voting Classifier)

ABC Adaptive Boosting (La methode de stimulation)

BC Bagging Classifier (La methode d'ensachage)

Chapitre 1

Introduction Général

1.1 Contexte

Le marché de télécommunication se développe de jour en jour. Les entreprises font face à une perte de revenus importante en raison de la concurrence croissante, d'où la perte de clients. Ils essaient de trouver les raisons de perdre des clients en mesurant la fidélité des clients pour retrouver les clients perdus.

La prédiction du taux de churn (désabonnements des clients) des clients dans l'industrie des télécommunications est un sujet de recherche centrale au cours des dernières années. Une énorme quantité de données sont générées dans ce secteur chaque minute. D'un autre côté, il y a beaucoup de développement dans les techniques d'exploration de données connue par le data Mining ou machine Learning.

Le taux de churn des clients est devenu l'un des principaux problèmes de l'industrie des télécommunications. Le taux mensuel moyen de churn chez les opérateurs de téléphonie mobile en Europe varie entre 8 et 12% [1], le taux d'attrition annuel varie de 20% à 40% chez la plupart des opérateurs de téléphonie mobile [2].

Dans un marché d'une telle compétitivité, une stratégie de marketing défensive revêt beaucoup l'importance. Au lieu de tenter d'acquérir de nouveaux clients ou d'attirer les abonnés loin de la concurrence, le marketing défensif s'intéresse plutôt à la réduction des départs de ses clients [2], surtout qu'il est 5 fois plus coûteux d'acquérir un nouveau

client que d'en garder un [3]. Et afin de fidéliser les clients, les fournisseurs de télécommunications doivent connaître les raisons du taux de désabonnement, qui peuvent être réalisées grâce aux connaissances extraites des données Télécom.

1.2 La problématique

Dans le passé, le churn a été identifié comme un problème dans la plupart des secteurs industriels.

Dans son sens le plus général, il se réfère au taux de perte des clients de l'entreprise. Il y a une raison simple pour le churn de l'attention attirée : perte des clients signifie une perte de revenus. Émergeant d'espaces d'affaires comme les prestataires de services de covoiturage, où le churn est un problème majeur, d'où la nécessité de s'attaquer à ce problème.

1.3 Les objectives

L'objectif de ce projet est de comparer des modèles de prédiction du churn dans le secteur de télécommunication et choisir les meilleurs entre eux.

1.4 La structure de mémoire

Pour mener à bien nos recherches, nous avons structuré nos travaux en quatre chapitres :

Le premier chapitre est consacré à la gestion électronique des relations avec la clientèle, y compris les notions générales sur la gestion des relations avec la clientèle, puis nous avons présenté une définition profonde de churn et ses types.

Dans le deuxième chapitre nous avons décrit la nécessité de l'apprentissage machine pour la détection de churn, et la manière générale dont il fonctionne pour ce faire. Ensuite, les différentes approches d'évaluation des modèles d'apprentissage automatique pour optimiser la précision de la détection.

Le troisième chapitre visé comprend toutes les étapes concernant les données, le prétraitement, les analyses nécessaires, afin de rendre les données exploitables et plus significantes, pour former les algorithmes d'apprentissage automatique.

Pour le quatrième et le dernier chapitre nous avons mené la mise en œuvre de notre système à travers les techniques adéquates et en utilisant le matériel disponible, ensuite nous avons présenté les résultats et les tests de notre solution ainsi que l'évaluation des modèles.

Chapitre 2

Clients Churn

2.1 Introduction

Parlez à tous ceux qui participent au marketing dans l'industrie des télécommunications aujourd'hui, et les chances sont bonnes que votre discussion va tournera rapidement vers la question du churn. Le churn, semble-t-il, est un gros problème pour de nombreuses industries des télécommunications, et pour une bonne raison.

Ce chapitre comporte trois sections essentielles :

La première section contient les définitions de la gestion de la relation client CRM et la gestion électronique de la relation client l'E_CRM, et quelles sont les différences entre eux.

La satisfaction du client est le premier principe mis en avant par la norme ISO 9001, avant que l'organisme commence à identifier les besoins et les attentes de ses clients, la deuxième section II explique leur importance.

Le churn est le terme qui a été adopté pour définir le mouvement des clients d'un fournisseur à un autre. La troisième et la dernière section présentent une définition profonde de churn.

2.2 Gestion de la Relation Client CRM

2.2.1 Définition et Principes de CRM

CRM signifie Customer Relationship Management en anglais soit littéralement Gestion de la Relation Client (GRC). Selon les organisations, cette approche peut également être désignée par la notion de Marketing client, par opposition au Marketing produit.

La gestion de la relation Client est une stratégie business par laquelle l'entreprise vise à comprendre, gérer et anticiper les besoins de ses clients actuels et potentiels. Cette gestion repose sur le principe de fidéliser ses clients que d'en acquérir de nouveaux. Il est communément admis que le coût d'acquisition d'un nouveau client est en moyenne entre 3 à 5 fois supérieur aux investissements consentis pour conserver ceux déjà acquis.

Cette approche se justifie également par la « théorie de la baignoire » qui trouve tout son sens dans les contextes fortement concurrentiels et/ou de contention de budgets marketing : recruter de nouveaux clients est nécessaire, notamment en phase de croissance et de développement (nouvelle entité, nouvelle offre etc.) : c'est la phase de conquête extensive.

Mais si la baignoire n'est pas bouchée, les flux de nouveaux clients entrent dans le portefeuille de l'entreprise et en ressortent naturellement. Cette situation devient un sujet de préoccupation lorsque le nouveau client est devenu Ancien avant même d'avoir permis à la marque de rentabiliser les investissements opérés pour le recruter (coûts fixes de communication media notamment).

La gestion de la relation client (GRC ou CRM) désigne l'ensemble de la démarche qui, à partir d'une base de données et d'applications logicielles spécifiques, permet de pratiquer un marketing ouvert (multiplier les points de contact) et relationnel avec ses clients dans le but d'augmenter la rentabilité globale de l'entreprise. La gestion de la relation client consiste à savoir cibler, à attirer et à conserver les bons clients et représente un facteur déterminant du succès de l'entreprise.

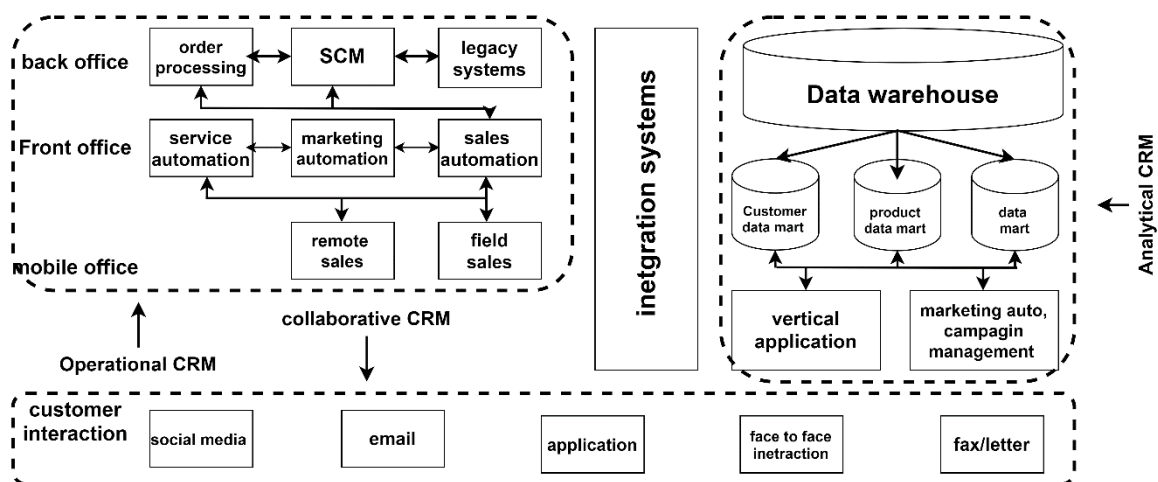
Construire et développer des relations avec ses clients est un challenge, particulièrement lorsque l'entreprise possède des milliers (voir des millions) de clients qui communiquent avec celle-ci de multiples manières. Pour arriver à un résultat satisfaisant, les

systèmes de gestion des relations clients doivent permettre aux responsables d'entreprise de mieux comprendre leurs clients pour adapter et personnaliser leurs produits ou leurs services.[4]

2.2.2 Les Trois Parties du CRM

La description complète du fonctionnement du CRM dans une entreprise serait trop complexe. Certains auteurs s'intéressent à une vue de (Dohnal, J., 2002) qui décrit cela comme les trois parties de l'architecture CRM, Alors que (Buttle, F., 2009) appelle cette architecture par « types de CRM », selon (Dohnal, J., 2002) Le CRM peut être organisé en trois grands domaines : opérationnel, analytique et collaboratif. [5]

FIGURE 2.1 – Le modèle de système de CRM qui comprend tous les types



Source : JELINEK (D) "The Evolution of Customer Relationship Management System."

Proceedings of the 19th International Conference on Computers, 2015, p.30.[6]

Le CRM Opérationnel

(Bose, R., 2002) a souligné que le CRM est l'intégration des technologies et des processus d'affaires, qui sont adoptés pour satisfaire les besoins d'un client au cours d'une interaction donnée.

Le CRM opérationnel est centré sur la gestion quotidienne de la relation avec le client, à travers l'ensemble des points de contact; centres de contacts à distance, téléphone, internet et outils de force de vente. Ce domaine implique l'automatisation des processus

qui touchent les départements en contact avec les clients : commercial, marketing et service client, via les différents canaux d'interaction.[7]

Le CRM Analytique

Mariage de la relation client et de l'analyse des données. Ce domaine s'appuie sur les entrepôts de données qui regroupent les données provenant des systèmes de CRM opérationnel et des points de contact avec les clients et les soumet à des analyses, il est intimement lié au data warehouse (entrepôt de données) et aux applications décisionnelles. Le CRM analytique est le levier du retour sur investissement des projets du CRM. Sans outil décisionnel associé, un progiciel de CRM ne sert qu'à stocker des données sans offrir de réelles capacités à les analyser.[7]

Le CRM Collaboratif

Le CRM se traduit par la mise en œuvre de techniques collaboratives destinées à faciliter la communication entre l'entreprise et ses clients ainsi que l'intégration avec les autres départements de l'entreprise : logistique, finances, production et distribution.[7]

2.2.3 Les Objectifs de CRM

L'objet du CRM est d'être plus à l'écoute du client afin de répondre à ses besoins et de le fidéliser. Un projet CRM consiste donc à permettre à chaque secteur de l'entreprise d'accéder au système d'information pour être en mesure d'améliorer la connaissance du client et de lui fournir des produits ou services répondant au mieux à ses attentes. Le CRM (et l'E_CRM plus particulièrement) est une stratégie qui sert de moyen pour se rapprocher des clients afin de les satisfaire au mieux et ainsi gagner en croissance et en rentabilité.[4]

Acquérir de nouveaux clients

Le préalable pour recruter de nouveaux clients est d'obtenir les adresses des futurs prospects (clients potentiels). Il est possible d'acheter des listes d'adresses auprès

d'entreprises spécialisées. La collecte peut également se faire par les canaux CRM offline (coupon-réponse papier,) ou online (via des formulaires de contact, jeux concours, inscription à la newsletter, transactions électroniques, demandes de téléchargement, co-abonnement, ...).

Ensuite il s'agit d'attirer les prospects en leur proposant une offre attractive : par exemple abonnement à un service, adhésion à un club ou souscription à une opération. Le nombre de prospects qui ont été transformés en clients par rapport aux messages émis (taux de transformation) permet de connaître le rendement d'une opération de marketing et le coût d'acquisition d'un client. Pour une entreprise, acquérir de nouveaux clients demande beaucoup d'investissements pour des résultats parfois médiocres. Les spécialistes estiment qu'il est nettement plus facile de fidéliser les clients existants.

Connaître le client

Pour mieux connaître ses clients, une entreprise doit rassembler les informations reçues (les coordonnées du client, ses préférences en termes de produit et de service, l'historique de ses achats, messages échangés, envois et réceptions,...) lui permettant de décrire et de caractériser sa clientèle.

Ces données, qui peuvent être massives, sont stockées dans des entrepôts de données (Data Warehouse) orientés clients. Celles-ci sont ensuite analysées. La technique d'analyse la plus utilisée est le datamining.

Dans le domaine de l'E_CRM et du Web, le datamining est souvent utilisé pour classer les internautes en segments d'utilisateurs, définir les profils de navigation et établir des liens entre les produits.

Ainsi le datamining permet le profilage des clients en catégories afin de leur faire des offres qui leur sont adaptées. Plus une offre est adaptée au client auquel elle s'adresse, plus il y a de chances que ce client soit intéressé à l'acheter.

Fidéliser les clients existants

L'acquisition d'un client par les différentes opérations de marketing représente un coût élevé pour l'entreprise. Il faut donc que le client acquis soit rentable. Pour cela, le

cycle de vie d'un client doit être prolongé au maximum. Or puisque l'internet est devenu très concurrentiel (pratiquement chaque entreprise possède au moins un site internet) et que les clients sont volatiles, rien n'empêche les clients même très satisfaits d'acheter chez des concurrents. Un client n'est donc jamais définitivement acquis. Il faut alors l'amener à acheter régulièrement et même à augmenter son panier d'achat.

Le client recruté ne doit absolument pas devenir un client inactif (qui n'a acheté qu'une seule fois), sinon l'entreprise y perd. Celle-ci doit donc tout mettre en œuvre pour garder l'attention (fidéliser) de ses clients et les amener à consommer ses produits et services régulièrement par des offres adaptées.

2.2.4 Définition de E_CRM

Les progrès de la technologie, de l'informatique et des télécommunications ont soutenu le développement de la technologie Internet. L'avancement rapide de la technologie dans tous les secteurs a également encouragé le développement progressif du concept de CRM, conduisant à l'émergence du concept électronique de gestion de la relation client (E_CRM).[8]

Le concept E_CRM désigne l'ensemble des activités, outils et techniques marketing qui permettent de construire et d'améliorer la relation entre l'entreprise et ses clients et de développer des relations avec les clients existants et potentiels via Internet, il peut être décrit comme l'utilisation de la technologie pour soutenir la stratégie de gestion de la relation client et est considéré comme l'intégration de la gestion de la relation client avec les applications e-business.

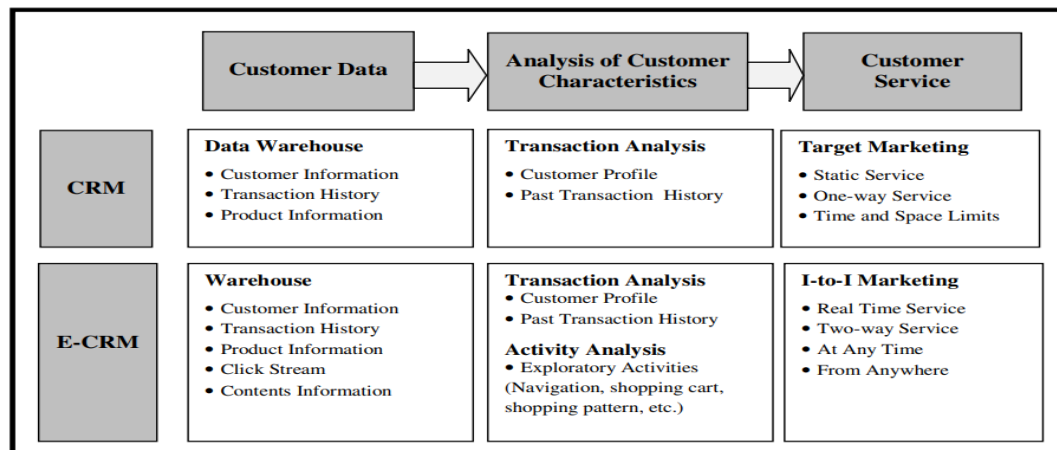
Le développement du marketing par le biais des médias en ligne est en augmentation dans laquelle E_CRM aide à gérer la relation avec les clients et fournit plusieurs avantages pour les entreprises comme l'augmentation des bénéfices des entreprises, créer des services satisfaisants en utilisant des informations intégrées et en démontrant la cohérence, la procédure et le processus de traitement des problèmes. [9]

2.2.5 La Différence entre CRM et E_CRM

Les différences entre CRM et E_CRM sont subtiles mais essentielles, elles concernent la technologie sous-jacente et ses interfaces avec les utilisateurs et d'autres systèmes. Par exemple, E_CRM offre la possibilité de prendre soin des clients via le Web, ou les clients étant en mesure de prendre soin d'eux-mêmes en ligne. De nombreux systèmes E_CRM fournissent au client une fenêtre libre-service basée sur le navigateur pour passer des commandes, vérifier l'état des commandes, examiner l'historique des achats, demander des informations supplémentaires sur les produits, envoyer des e-mails et s'engager dans une masse d'autres activités.[10]

Le client n'est plus limité à communiquer avec une organisation pendant les heures normales d'ouverture, et l'organisation n'a pas à fournir une personne-ressource à l'autre bout pour les demandes de renseignements et les demandes des clients. En effet, dans un environnement E_CRM, les clients font la plupart du travail pour eux-mêmes et non pour les entreprises. [11]

FIGURE 2.2 – La différence entre CRM et E_CRM



Source : PAN (S. L), LEE (J. N) : « Using E_CRM for a unified view of the customer,»

Communications of the ACM, 46(4), 2004, p.95–96.[11]

2.3 Satisfaction de la clientèle

2.3.1 Le Client : la personne la plus convoitée

Jusqu'à la fin des années 1980, la grande majorité des distributeurs faisait la même chose. Ils mettaient à la disposition des consommateurs des masses de produits : la demande était supérieure à l'offre ! À cet égard, le terme « distribution » énonce clairement la vocation initiale de ce type de commerce.

Dans l'intervalle, la succession de crises économiques et la baisse du pouvoir d'achat, mais aussi certains abus tels que des défauts de qualité, ont altéré le capital confiance des consommateurs. Le client a évolué : il est devenu plus adulte devant les tentations du libre-service.

Le mouvement s'est initié dans les années soixante aux États-Unis, où les liges de défense du consommateur menées par Ralph Nader ont fait naître un nouvel état d'esprit : « on ne nous fera pas consommer n'importe quoi ! » Ce nouveau comportement de « vigilance » est aujourd'hui consacré : le consommateur est devenu consumériste.

Rappelons-nous : au début du libre-service, faire ses achats était pour le consommateur un plaisir (qui est maintenant devenu une corvée) ! Ainsi, les consommateurs, à la recherche de prix bas, sont aussi prêts à payer « pour se faire plaisir » des produits ou des services à valeur ajoutée. Comme l'écrit Stéphane Sabbah dans un éditorial de la revue *Échanges* de juin 2007, en ouverture d'un excellent dossier consacré aux nouvelles tendances de la distribution : « confrontées à des nouvelles législations (L. Dutreil, une concurrence accrue hard discount) et à un pouvoir d'achat atone, les enseignes ont dû développer de nouvelles portes de croissance : fidélisation des clients via les cartes de fidélité, marque de distributeurs, distribution via Internet... »

Dans ce contexte à la recherche d'un nouveau modèle de croissance, se dessinent des démarches au cœur desquelles le client, objet de toutes les convoitises, conduit la grande distribution à passer d'une entreprise logistique à une entreprise à dominante marketing.

Face à ces évolutions, les distributeurs ont dû repositionner leur stratégie d'enseigne afin de créer la différence, fidéliser et attirer de nouveaux consommateurs.

Chaque enseigne a donc effectué un travail de fond, afin de renforcer son originalité dans

sa politique d'assortiment, de prix, de service, d'implantation, de présentation, etc.[12]

2.3.2 Les Cinq pouvoirs du client

Le client a bien changé : de consommateur, il est devenu consommériste. Aujourd'hui, il a conscience de son pouvoir, il est plus informé donc plus critique vis-à-vis de l'offre. Il n'hésite pas à sanctionner positivement ou négativement en choisissant de donner ses euros à l'enseigne qui le considère le plus !

Face à la pléthore d'offres, le client adopte un comportement plus avisé et n'accepte de se laisser séduire que si l'enseigne répond à ses attentes.[12] Ces nouveaux pouvoirs reposent sur cinq leviers :

- Le choix auquel est confronté le client.
- L'information à laquelle il a accès.
- Le poids de ses dépenses.
- Le temps dont il dispose pour consommer.
- L'influence qu'il peut exercer sur la marque ou l'enseigne.

2.3.3 Satisfaction de la clientèle

Avant qu'une organisation n'évalue la satisfaction de ses clients, il faut qu'elle la définisse. Plusieurs définitions ont été données pour la satisfaction client et chacune repose sur des critères et facteurs spécifiques.

La recherche en sciences humaines montre cependant que ces facteurs et leurs impacts ne sont pas bien précisés ou définis. Par exemple, la relation Longue terme entre le client et le fournisseur est considérée comme une signification importante de la satisfaction du client. Au contraire, d'autres travaux scientifiques voient que la relation long-terme ne donne aucune signification sur la satisfaction du client car cette relation peut être issue d'un engagement, d'une habitude, d'une confiance, etc.

La satisfaction du client peut également être reflétée par le jugement, l'avis, l'opinion du client sur un produit acheté ou un service rendu. Il s'agit alors d'une évaluation « après-achat » ou après consommation du service/produit. Afin de mesurer la satisfaction

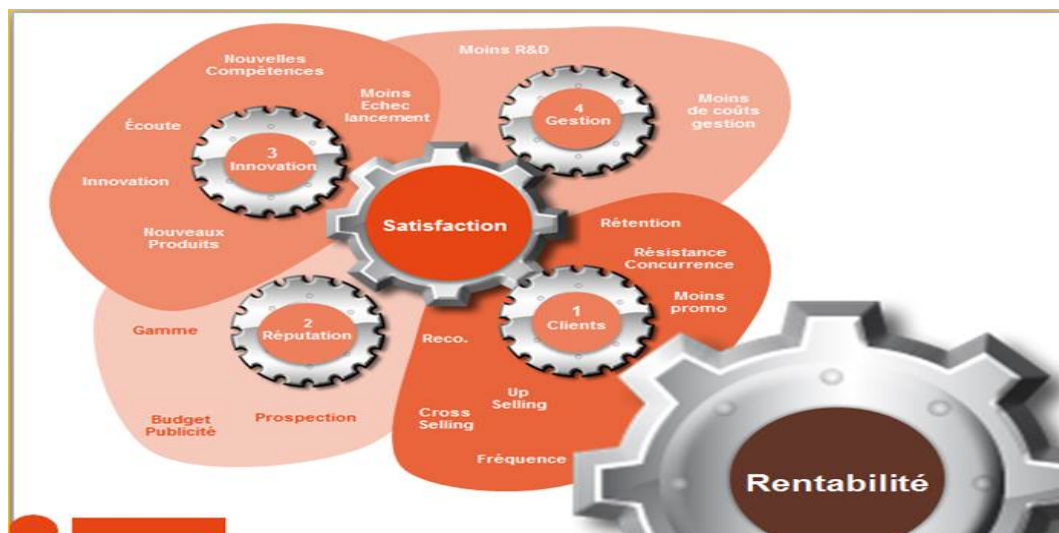
client, une comparaison peut être faite entre la performance du service/produit perçu par le client avec un standard préétabli. Cette comparaison donne dès lors les écarts et les améliorations possibles pour satisfaire le client.[13]

2.3.4 Les 19 Impacts de la satisfaction client dans la rentabilité des entreprises

La recherche de la satisfaction client est un moteur extrêmement puissant dans l'optimisation de la rentabilité des entreprises. Les dirigeants qui ont la volonté d'intégrer la satisfaction client comme un élément de management et de motivation de leurs équipes ont compris des choses que les autres ignorent !

Le bon taux de satisfaction client n'est pas la raison du succès, ce n'est que la conséquence de la volonté d'un manager qui a su en faire une valeur dans son entreprise et qui a créé un écosystème et une culture d'entreprise qui font la différence. Ces managers ont implicitement intégré le cycle vertueux des 19 rouages qui conduisent de la satisfaction client à la rentabilité![14]

FIGURE 2.3 – Les 19 impacts de la satisfaction client



Source : <https://www.paperblog.fr/7632162/les-19-impacts-de-la-satisfaction-clients-sur-la-rentabilite-des-entreprises/>. [15]

Des clients satisfaits

1. Sont moins sensibles aux promotions.
2. Achètent plus le même produit (up selling) plus souvent ou avec des options.
3. Achètent plus souvent d'autres produits (cross selling).
4. Résistent plus facilement à la pression de la concurrence.
5. Sont moins infidèles (impact sur la gestion).
6. Recommandent Entreprise a leur entourage (impact sur la réputation).

Impact sur la réputation

7. Des clients satisfaits contribuent à l'image de marque et à la bonne réputation d'entreprise.
8. Cela favorise la prospection et le recrutement des nouveaux clients via un bouche-à-oreille favorable.
9. Cela optimise les investissements commerciaux et publicitaires.

Impact sur L'innovation

Une entreprise qui a la volonté de satisfaire : :

10. Est une entreprise qui sait écouter.
11. Est plus tournée vers innovation produits et services.
12. Elle peut acquérir de nouvelles compétences.
13. Optimise sa politique de R&D (rechercher et développer) fondée sur les vraies attentes des clients.
14. Connait moins d'échecs lors des lancements de nouveaux produits.
15. Mets de « bons produits » sur le marché.
16. Elargit sa gamme (ce qui impacte son image).

Impact sur la gestion

17. Gérer et facturer 1 000 € à un client, est plus rentable que de gérer et de facturer 100 € a 10 clients.

18. Optimisation d'efficacité des investissements commerciaux et publicitaires via le bouche-à-oreille et une meilleure réputation.
19. Optimisation du potentiel offert par chaque client.

2.4 Le Churn du Client

2.4.1 Valeur à vie du client (Customer Lifetime Value CLV)

Le CLV est défini comme «la valeur actuelle des bénéfices futurs générés par un client au cours de sa vie d'affaires avec l'entreprise ». Il tient donc compte des revenus et des coûts.

Une clientèle plus fidèle peut mener à un CLV plus élevé grâce à une plus grande volonté de payer, une plus grande part de portefeuille, une relation plus longue avec l'entreprise ou des ventes croisées et ascendantes. Les clients fidèles affichent habituellement, mais pas nécessairement, un CLV plus élevé.

L'importance du CLV pour la prise de décisions de gestion est appuyée par plusieurs études. Par exemple, Gupta, Lehmann et Stuart (2004) confirment le lien positif entre la CLV et la valeur de l'entreprise, tandis que Kumar et ses collègues (2008) montrent que l'optimisation de l'affectation des ressources en fonction de la CLV peut augmenter les revenus.[16]

2.4.2 Définition de Churn

Selon Sharma et Panigrahi (2011), churning désigne un client qui quitte une entreprise pour aller dans une autre entreprise, ce qui entraîne non seulement une perte de revenu, mais aussi d'autres effets négatifs sur l'exploitation des entreprises (Chen et coll., 2014). Comme Hadden et al. (2005) l'ont stipulé, «la gestion du churn est le concept qui consiste à identifier les clients qui ont l'intention de transférer leur coutume à un fournisseur de services concurrents. ».

En ce qui concerne l'industrie des télécommunications, les clients quittent l'entreprise actuelle et passent à une autre entreprise de télécommunications. Avec le nombre croissant

de barattes, il devient le processus de l'opérateur de conserver les clients rentables connus sous le nom de gestion du churn.

Dans l'industrie des télécommunications, chaque entreprise offre aux clients d'énormes incitations pour les inciter à passer à leurs services, c'est l'une des raisons pour lesquelles le churn des clients est un gros problème dans l'industrie de nos jours. Pour éviter cela, l'entreprise doit connaître les raisons pour lesquelles les clients décident de passer à une autre entreprise de télécommunications. Il est très difficile de garder les clients intacts pendant une longue période, car ils bénéficient du service qui répond à la plupart de leurs besoins.[16]

2.4.3 Types de Churn

Sur la base de la méthode de révocation, les churns sont classés comme volontaires et involontaires.[17]

FIGURE 2.4 – Involontaires vs Volontaires Churn



Source : <https://blog.2checkout.com/keep-voluntary-churn-at-minimum/>. [18]

Le Churn Involontaires

Les churners involontaires sont les plus faciles à identifier. Ce sont les clients que Telco décide de supprimer de la liste des abonnés. Par conséquent, cette catégorie comprend les personnes qui sont victimes de fraude, de non-paiement et les clients qui n'utilisent pas le téléphone.

Le Churn Volontaires

Le churning volontaire est plus difficile à déterminer ; il se produit lorsqu'un client prend la décision de mettre fin à son service avec le fournisseur. Lorsque les gens pensent à Telco churning est généralement le genre volontaire qui vient à l'esprit.

Le churning volontaire peut être subdivisé en deux catégories principales, le churning accidentel et le churning délibéré.[19]

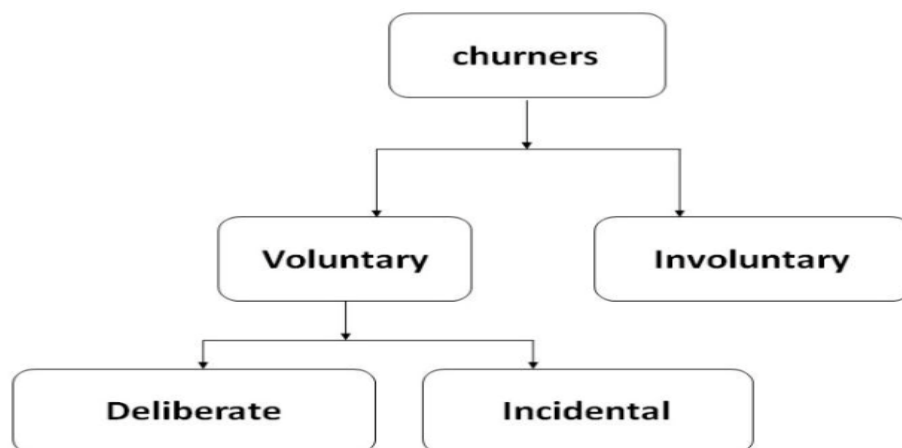
1. Le Churn Accidentel

Le churning accidentel se produit, non pas parce que les clients l'ont prévu, mais parce que quelque chose est arrivé dans leur vie. Par exemple : changement de la situation financière, changement de l'emplacement, etc.

2. Le Churn Délibéré

Le churning délibéré pour des raisons de technologie (clients désirant une technique plus récente ou meilleure), d'économie (sensibilité aux prix), de qualité de service, de facteurs sociaux ou psychologiques, et de commodité. Le churning délibéré est le problème que la plupart des solutions de gestion du churning tentent de résoudre.

FIGURE 2.5 – La taxonomie de churning



Source : <https://www.semanticscholar.org/paper/Modeling-%26-Simulation-of-a-Predictive-Customer-for-AwoyeluI./6f41c3d1dbd05567fa8b3e455bad5343647e94ed>. [20]

En ce qui concerne le contexte commercial, le churning peut être catégorisé en grande partie comme le churning contractuel et le churning non contractuel.[21]

Le Churn Contractuel

Dans un churn contractuel, les clients prendraient certaines mesures à des intervalles discrets. Dans ce type de churn, les révocations sont observées explicitement.

Le Churn Non Contractuel

Le churn non contractuel se produit lorsque les clients cessent d'adopter une certaine conduite au fil du temps. Cela décrit les circonstances où le churn ne peut pas être détecté rapidement sur une action spécifique du client. Ainsi, ce type de churn est abstruse en termes de compréhension. Exemple de ce churn comprend un détaillant de mode en ligne parce qu'ici les clients sont libres d'acheter ou pas à tout moment.

2.4.4 Les Déterminants du Churn

Statut de client

Certains clients n'abandonnent pas soudainement un fournisseur de services. En fait, ils décident de ne pas l'utiliser de façon intempestive ou sont suspendus par le fournisseur de services en raison de problèmes de paiement, c'est-à-dire qu'ils présentent un changement de statut dans une base de données interne sur les clients de l'entreprise. Aux fins de la présente étude, plus d'un centième de types différents de statut de client sont regroupés en trois catégories : utilisation active, non-utilisation et en suspension. Les clients dont le statut n'est pas d'utilisation ou est suspendu sont considérés comme plus susceptibles de basculer que les clients dont le statut est d'utilisation active.[22]

Le prix du service

Le prix du service désigne le montant d'argent que le client devrait payer pour recevoir des services. En fait, les clients cherchent des marchés avec moins de prix de service. En conséquence, en raison de ce fait que les clients sont attirés par les concurrents avec des prix plus bas, les fournisseurs de services essaient de réduire leurs coûts pour attirer plus de clients et moins de pertes de la clientèle. Ainsi, on peut soutenir que des prix plus élevés ont des effets négatifs sur l'achat des clients et des effets positifs sur La perte des clients.

Selon diverses recherches, la perception d'un prix équitable est un facteur influent sur le succès des entreprises. Lorsque les clients comparent le prix attendu avec le prix qu'ils ont payé. Si un prix est perçu comme étant inférieur au prix de référence, ils perçoivent l'équité du prix et la valeur de la transaction, sinon ils ont l'impression de perdre de la valeur.[22]

Le coût de commutation (Switching cost)

Le coût de commutation est défini comme le coût encouru lorsqu'il interdit aux clients de passer au service des concurrents. Bien sûr, ces coûts ne sont pas seulement économiques, mais aussi physiques, émotionnels et temporels.

En fait, lorsque les clients passent à des concurrents, ils perdent surtout du temps, de l'énergie et de l'argent, même s'ils peuvent être exclus de certains avantages et occasions particulières en raison d'être membre d'une organisation spécifique.

Par conséquent, le changement de fournisseur de services entraîne divers coûts et exclut même les clients de certains avantages. Ainsi, si les clients font face à des coûts de commutation élevés, ils préfèrent ne pas changer de fournisseur de services pendant qu'ils sont insatisfaits. On peut en conclure que le coût de commutation a des effets négatifs sur la perte des clients.[22]

La qualité

La qualité est la différence entre percevoir ce que les clients attendent et ce qu'ils perçoivent. Grönroos (1997) a déclaré que la qualité du service est la différence entre les services réels et les services promis. Par conséquent, dans les organisations de service, la qualité dépend de la quantité d'attentes des clients répondues par l'organisation.

Dans les entreprises de services mobiles, la qualité du service fait référence à la qualité des appels, soit les services audios, vidéo et texte fournis par l'opérateur pendant un appel. Ainsi, la qualité peut être mentionnée comme l'un des facteurs influents sur le désabonnement de la clientèle.[22]

2.4.5 Le Churn est Coûteux

Non seulement le churn est un parti inévitable de faire des affaires dans le sans-fil, il s'avère également être une expérience extrêmement coûteuse. Le churn a de nombreuses conséquences, et la plupart d'entre elles ont un prix élevé.

La plus grande conséquence du churn est, bien sûr, la perte de revenus. Selon le pays, le client moyen rapporte entre 20 \$ US et 80 \$ US par mois. La perte d'un grand nombre de clients peut créer une grande brèche dans le bilan de l'entreprise.

La perte directe de revenus par la réduction du nombre d'abonnés n'est qu'un des coûts réels associés au churn. La réduction parfois drastique des taux de facturation est également endémique à l'environnement de churn.

Lorsque le churn commence à se produire, la plupart des entreprises réagissent en baissant leurs prix. Ils essaient de convaincre les clients qu'ils sont compétitifs et que les clients n'ont pas besoin de partir pour obtenir de bons prix. Cette réduction des taux entraîne logiquement une réduction des revenus annuels de l'entreprise.[16]

Le Coût de la Réacquisition

Malgré tous leurs efforts pour empêcher le churn, l'entreprise perdra inévitablement certains de ses clients à la concurrence tôt ou tard.

Lorsqu'ils le font, ils se rendent souvent compte qu'il est possible de les reconquérir en menant des campagnes de réacquisition. Ces campagnes sont souvent couronnées de succès mais, bien sûr, entraînent leurs propres coûts.

Le Coût de la Fidélisation de la Clientèle

Les organisations plus proactives créent en fait des campagnes pour aider à empêcher les gens de partir. Ces campagnes de fidélité ou de rétention permettent à l'entreprise de faire des offres ou de créer des expériences de relation client qui aident le consommateur à décider de ne pas quitter le fournisseur actuel.[23]

2.4.6 Le Churn est difficile à Gérer

Vous pouvez trouver des problèmes de churn partout dans le monde, mais l'omniprésence du churn n'est pas ce qui le rend si intéressant pour les gens. Non, la qualité la plus déconcertante du churn est qu'il est très difficile à gérer. Pour de nombreuses raisons, le churn peut rapidement devenir le plus gros problème auquel l'organisation est confrontée.[23]

La Grosse surprise

La chose la plus ennuyeuse au sujet des problèmes de churn est qu'ils semblent venir soudainement et sans attente. À maintes reprises, des dirigeants nous ont raconté comment ils s'en sortaient, comment ils faisaient les choses comme d'habitude, quand un jour ils se sont rendu compte que le taux de churn de la clientèle avait en quelque sorte dérapé.

Il y a de nombreuses raisons pour lesquelles le churn surprend les gens, certaines évidentes, d'autres pas si évidentes. À l'heure actuelle, cependant, suffisamment de preuves ont été rassemblées pour que les gens se rendent compte qu'ils devraient planifier pour que cela se produise tôt ou tard.

Difficile de prédire

Non seulement il est difficile d'anticiper le churn, mais il est aussi difficile de prédire l'ampleur du problème, sa durée et les conséquences du phénomène. Il y a un certain caractère aléatoire à baratter qui peut être très intimidant pour les gens, au moins jusqu'à ce qu'ils apprennent à mieux le comprendre.

Difficile à expliquer

Lorsque votre entreprise commence à éprouver le churn, l'une des expériences les plus exaspérantes sera de comprendre pourquoi il se produit. Après tout, si vous saviez pourquoi les clients partaient, vous pourriez faire quelque chose. Malheureusement, la plupart des entreprises sont mal préparées à donner ce genre d'explication.

Difficile à défendre

Et, bien sûr, si vous ne savez pas quand il se produira, ou pourquoi, alors vous n'aurez pas beaucoup de chance de mettre en place un plan qui vous permet de défendre contre la perte de vos clients.

2.4.7 Churn vous aide à apprendre à propos de l'entreprise

Lorsque les dirigeants d'entreprises sont confrontés à des problèmes de churn pour la première fois, cela peut être une expérience des plus démoralisantes. Tout à coup, la société se rend compte qu'elle n'a aucune idée de qui sont ses clients ou de ce qu'ils veulent.[16]

Découvrir qui sont vos clients

Le churn est une décision incroyablement complexe et personnelle pour la plupart des gens. Cela signifie que, pour vraiment gérer le churn de manière efficace, nous devons savoir exactement qui sont nos clients.

Les stratégies de gestion du churn qui traitent tous les clients comme essentiellement les mêmes, donnent de très mauvais résultats et, donnent lieu à des revenus qui n'ont pas besoin d'être perdus.

Apprendre ce que nos clients veulent

En plus de déterminer qui sont nos clients, nous devons trouver un moyen de déterminer exactement ce que ces clients veulent et ce dont ils ont besoin.

La survie de l'entreprise de covoiturage de l'avenir dépendra non seulement de leur capacité à comprendre qui sont leurs clients et ce qu'ils veulent aujourd'hui, mais aussi d'anticiper les besoins que les consommateurs auront à l'avenir rapidement, précisément, économiquement, et avant que la concurrence ne le découvre.

Apprendre ce que notre entreprise devrait faire ensuite

Cela nous amène à la vraie leçon que le churn a pour nous et notre organisation. Le churn peut nous apprendre comment gérer notre entreprise différemment. Il peut nous apprendre à le faire fonctionner, non pas en fonction des exigences de la technique, mais en fonction des besoins des consommateurs. Apprendre à anticiper les besoins des clients et construire une organisation capable de répondre rapidement à ces changements de besoins est les clés de votre succès à l'avenir.

2.5 Conclusion

Dans ce chapitre, nous avons appris que l'expression de la gestion de la relation client CRM a une variété de significations, trois types de CRM ont été identifiés : opérationnel, analytique et collaboratif, et nous avons discuté la gestion électronique de la relation client l'E_CRM et sa différence avec CRM.

Nous avons présenté la satisfaction des clients et ses impacts dans l'organisation. Et abordé explicitement la définition du notre problème et nous nous sommes concentrés sur l'importance de cette étude sur le churn et sur la façon dont elle devrait être une préoccupation constante de toute entreprise aujourd'hui, peu importe le domaine de l'entreprise.

Dans le chapitre suivant, nous discuterons du rôle des techniques de machine learning dans le problème de churn.

Chapitre 3

L'Apprentissage Automatique (Machine Learning)

3.1 Introduction

Après avoir pris une idée sur le churn et ses concepts, nous allons dans ce chapitre nous intéresser en particulier à la partie "L'apprentissage automatique ML" qui doit être couverte pour représenter les outils et le périmètre sur lesquels notre projet est basé.

Nous commencerons par décrire la nécessité de l'apprentissage machine pour la détection de churn, et la manière générale dont il fonctionne pour ce faire.

Ensuite, nous présenterons les différentes approches de détection d'état de client (rester / quitter), basées sur l'apprentissage automatique, et nous conclurons par l'évaluation des modèles d'apprentissage automatique pour optimiser la précision de la détection.

3.2 L'Apprentissage Automatique

En 2006, Geoffrey Hinton et al. ont publié un article montrant comment former un réseau neuronal profond capable de reconnaître les chiffres manuscrits avec une précision de pointe (>98). Ils ont qualifié cette technique « D'apprentissage profond ». La formation d'un réseau neuronal profond était largement considérée comme impossible à

l'époque, et la plupart des chercheurs avaient abandonné l'idée depuis les années 1990. Cet article a ravivé l'intérêt de la communauté scientifique et, avant longtemps, de nombreux nouveaux articles ont démontré que l'apprentissage profond était non seulement possible, mais capable de réalisations époustouflantes qu'aucune autre technique d'apprentissage automatique (ML) ne pouvait espérer égaler (avec l'aide d'une puissance de calcul énorme et de grandes quantités de données). Cet enthousiasme s'est rapidement étendu à de nombreux autres domaines de l'apprentissage automatique.

Dix ans plus tard, le Machine Learning a conquis l'industrie : il est maintenant au cœur de la magie des produits high-tech d'aujourd'hui, en classant les résultats de vos recherches sur le web, en alimentant la reconnaissance vocale de votre smartphone et en recommandant des vidéos, battre le champion du monde au match de Go. Avant que vous le sachiez, il conduira votre voiture.[24]

3.2.1 Définition d'Apprentissage Automatique ML

L'apprentissage automatique est la science (et l'art) de la programmation des ordinateurs afin qu'ils puissent apprendre des données.

L'apprentissage automatique est une branche de l'intelligence artificielle. En utilisant l'informatique, nous concevons des systèmes qui peuvent apprendre des données de manière à être formés. Le système doit apprendre et s'améliorer avec l'expérience et, avec le temps, créer un modèle qui peut être utilisé pour prédire les résultats des questions en fonction de l'apprentissage précédent.

Ce terme fait référence à la capacité des systèmes de technologie de l'information (IT) de trouver indépendamment des solutions aux problèmes en reconnaissant les tendances dans les bases de données. En d'autres termes : la machine Learning permet aux systèmes informatiques de reconnaître des modèles sur les bases des algorithmes et des ensembles de données existantes et de développer des concepts de solutions adéquates.

Par conséquent, dans l'apprentissage automatique, les connaissances artificielles sont générées sur la base de l'expérience.[24]

3.2.2 L’Historique d’Apprentissage Automatique ML

Au cours des six dernières décennies, plusieurs pionniers de l’industrie ont travaillé pour nous orienter dans la bonne direction.[25]

Alan Turing

Dans son article de 1950 intitulé « Computing Machinery and Intelligence », Alan Turing demande : « Les machines peuvent-elles penser ? » Le document décrit le « jeu d’imitation », qui implique trois participants : un humain agissant comme juge, un autre humain et un ordinateur qui tente de convaincre le juge qu’il est humain. Le juge saisisrait un programme en phase terminale pour « parler » aux deux autres participants. L’humain et l’ordinateur réagiraient, et le juge déciderait quelle réponse venait de l’ordinateur. Si le juge ne pouvait pas toujours dire la différence entre les réponses humaines et informatiques, alors l’ordinateur a gagné le jeu.

Le test se poursuit aujourd’hui sous la forme du Prix Loebner, un concours annuel d’intelligence artificielle. Le but est assez simple : Convaincre le juge qu’ils discutent avec un humain au lieu d’un programme de chat bot de l’ordinateur.[25]

Arthur Samuel

En 1959, Arthur Samuel a qualifié l’apprentissage automatique de « domaine d’étude qui donne aux ordinateurs la capacité d’apprendre sans être explicitement programmés ». Samuel est crédité de créer l’un des programmes informatiques d’autoapprentissage avec son travail à IBM. Il s’est concentré sur les jeux comme un moyen d’obtenir l’ordinateur pour apprendre des choses. Le jeu de choix pour Samuel était des dames parce que c’est un jeu simple stratégie à partir de laquelle le programme pourrait apprendre. Avec l’utilisation de l’élagage d’évaluation alpha-bêta (éliminant les nœuds qui n’ont pas besoin d’évaluation) et minimax (minimisant la perte pour le pire cas) Samuel est largement connu pour son travail dans l’intelligence artistique, mais il a également été reconnu pour être l’un des premiers programmeurs à utiliser des tables de hachage, et il a certainement eu un grand impact chez IBM.[25]

Tom M. Mitchell

Tom M. Mitchell est le président de machine Learning à l'Université Carnegie Mellon. Comme auteur du livre Machine Learning (McGraw-Hill, 1997), sa définition de machine learning est souvent citée :

On dit qu'un programme informatique apprend de l'expérience E en ce qui concerne la classe de tâches T et la mesure de performance P, si sa performance atteint T, mesurée par P, s'améliore avec l'expérience E.

L'important ici est que vous avez maintenant un ensemble d'objets à définir l'apprentissage automatique :

- Tâche (T), soit une ou plusieurs
- Expérience (E)
- Performance (P)

Ainsi, avec un ordinateur exécutant un ensemble de tâches, l'expérience devrait mener à des augmentations de performance.[25]

3.2.3 Les Types d'Apprentissage Automatique

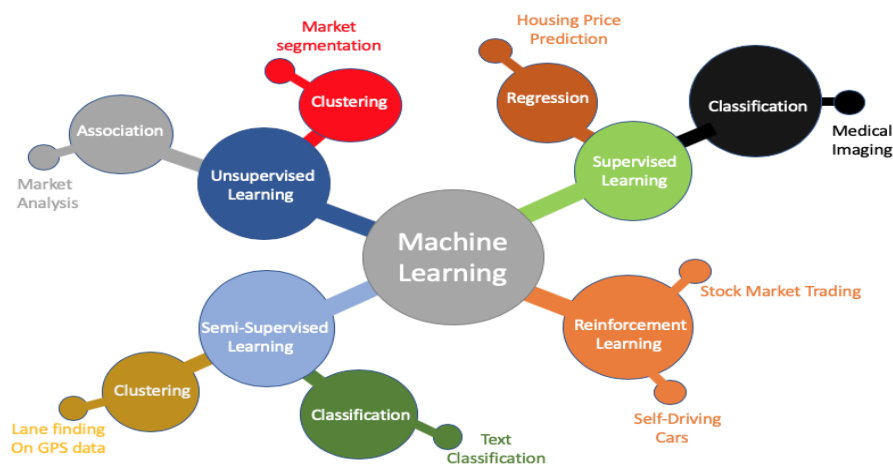
Il y a tellement de différents types de systèmes d'apprentissage automatique qu'il est utile de les classer dans de larges catégories basées sur :

- Qu'ils aient reçu ou non une formation sous supervision humaine (apprentissage assisté, semi-supervisé et renforcement)
- S'ils peuvent apprendre progressivement à la volée (en ligne ou par lots apprentissage)
- Qu'ils fonctionnent simplement en comparant de nouveaux points de données à des points de données connus, ou qu'ils détectent plutôt des tendances dans les données de formation et construisent un modèle prédictif, un peu comme le font les scientifiques (apprentissage basé sur des instances plutôt que sur des modèles)
- Qu'ils fonctionnent simplement en comparant de nouveaux points de données à des points de données connus, ou qu'ils détectent plutôt des tendances dans les données de formation et construisent un modèle prédictif, un peu comme le font les scientifiques (apprentissage basé sur des instances plutôt que sur des

modèles)

Les systèmes d'apprentissage automatique peuvent être classés en fonction de la quantité et du type de supervision qu'ils reçoivent pendant la formation. Il existe quatre grandes catégories : apprentissage supervisé, apprentissage non supervisé, apprentissage semi-supervisé et apprentissage par renforcement.[24]

FIGURE 3.1 – Les Types d'Apprentissage Automatique



Source :

<https://businesspartnermagazine.com/how-machine-learning-working-step-by-step/>. [25]

Apprentissage Supervisé

C'est le type d'apprentissage dans lequel l'algorithme apprend sur des données étiquetées. C'est-à-dire les données dont nous connaissons la sortie. Ici, un expert est employé pour étiqueter correctement des exemples.

Ces données d'apprentissage sont annotées. L'algorithme va donc produire une fonction de prédiction en analysant les similarités, ressemblances et caractéristiques qui distinguent les données ayant la même sortie. [26]

Apprentissage Non Supervisé

Contrairement à l'apprentissage supervisé, les données d'apprentissage ne sont pas étiquetées et que le nombre de classes et leur nature n'ont pas été prédéterminées. Il n'y a pas de résultats à prévoir. Au lieu de cela, l'algorithme d'apprentissage cherche une

structure dans ces données, en tirant des caractéristiques pour distinguer les exemples qui se ressemblent et les regrouper en classes.[27]

Apprentissage Semi-Supervisé

Ce type d'apprentissage utilise à la fois des données étiquetées et non étiquetées pour l'apprentissage. En général, une petite quantité de données annotées est utilisée. L'intérêt est de minimiser les ressources nécessaires pour analyser et tirer des informations des données étiquetées.[27]

Apprentissage par Renforcement

L'apprentissage par renforcement est caractérisé par la présence d'un agent qui doit apprendre un comportement par l'expérience, dans un environnement dynamique. C'est similaire à ce qui se fait en psychologie. Il est considéré comme une classe de problèmes au lieu d'un ensemble de techniques.[28]

3.2.4 La Nécessité de Machine Learning (ML) dans la Détection de Churn

Les machines sont bien meilleures que les humains pour traiter de grands ensembles de données. Elles sont capables de détecter et de reconnaître des milliers de modèles dans le parcours d'achat d'un utilisateur, au lieu des quelques modèles capturés par la création de règles.[29]

Les trois facteurs qui expliquent l'importance du ML sont les suivants :

Vitesse

Le ML peut évaluer un grand nombre de transactions en temps réel. Il analyse et traite en permanence de nouvelles données. De plus, un modèle avancé tel que les réseaux de neurones met à jour ses modèles de façon autonome pour refléter les dernières tendances.

Échelle

Les algorithmes et les modèles du ML deviennent plus efficaces avec l'augmentation des ensembles de données. L'apprentissage automatique s'améliore avec plus de données car le modèle ML peut mettre en évidence les différences et les similitudes entre plusieurs comportements.

Une fois qu'on leur a dit quelles les clientes fidèles, et les clients qui vont quitter, les systèmes peuvent les passer en revue et commencer à repérer celles qui correspondent à l'un ou l'autre de ces comportements. Ils peuvent également les prévoir à l'avenir lorsqu'ils traiteront de nouveaux clients.

Efficacité

Contrairement aux humains, les machines peuvent effectuer des tâches répétitives. De même, les algorithmes de ML font le gros du travail de l'analyse des données et ne transmettent les décisions aux humains que lorsque leur apport apporte un éclairage supplémentaire.

La découverte d'un modèle normal nécessite plus qu'un bon modèle ML. Une partie du processus de découverte de la normalité fait appel à l'intuition humaine, vous devez interagir avec le processus de modélisation pour décider de ce qui a un sens dans votre propre situation.

En construisant un modèle ML pour la détection d'anomalies, vous devez identifier le meilleur choix de données, trouver comment les mettre sous une forme acceptable pour votre algorithme et ensuite acquérir suffisamment de données pour entraîner votre modèle.

En d'autres termes, vous utiliserez les données dans un premier temps pour laisser le modèle découvrir des patrons que vous devrez ensuite interpréter afin de déterminer la situation de base ou normale.

Cela peut nécessiter un certain nombre d'ajustements de l'algorithme que vous utilisez avant de parvenir à quelque chose de logique. Plus vous en savez sur la situation étudiée, plus vous pouvez décider facilement et précisément quand votre modèle a atteint le premier objectif de détection des anomalies en trouvant ce qui est normal.

3.2.5 Fonctionnement d'un Système ML pour la Détection de Churn

L'image ci-dessous montre la structure de base du fonctionnement des algorithmes de détection de churn par apprentissage automatique :

FIGURE 3.2 – La structure de base du fonctionnement des algorithmes de détection de churn par l'apprentissage automatique



Source : Insights, R. Machine learning for fraud detection.[30]

Alimentation des Données

Tout d'abord, les données sont introduites dans le modèle. La précision du modèle dépend de la quantité de données sur lesquelles il est formé, plus il y a de données, plus le modèle est performant.

Pour détecter les clients qui vont désabonner (churn) spécifiques à une entreprise particulière, vous devez introduire de plus en plus de données dans votre modèle. Cela permettra de former votre modèle de manière à ce qu'il détecte parfaitement les activités de fraude spécifiques à votre entreprise.[31]

Extraction des Caractéristiques

L'extraction de caractéristiques consiste essentiellement à extraire les informations de chacun des fils associés à un processus de transaction. Il peut s'agir des données relatives aux services à la clientèle et aux informations contractuelles. En plus de tous les fournisseurs, forfaits et services auxquels le client est abonné. En outre, il contient

également des informations générées par le système CRM comme (type d'abonnement, anniversaire, sexe, le lieu de vie et plus ...).[31]

Entraînement de l'Algorithme

Une fois qu'un algorithme de prédiction de churn est créé, il doit être formé en fournissant des données sur les clients afin que l'algorithme de prédiction apprenne.[31]

Création d'un Modèle

Une fois que l'algorithme de prédiction de churn est formé sur un ensemble de données spécifiques, l'utilisation d'un modèle qui fonctionne pour détecter les clients "quitter" et "rester" dans l'entreprise est prêt.

L'avantage du ML des algorithmes de détection des churns est qu'il s'améliore constamment à mesure qu'il est exposé à de nouvelles données.[31]

3.3 L'Apprentissage Supervisé

L'apprentissage supervisé se réfère à une sorte de tâches d'apprentissage automatique axées sur la construction de modèles prédictifs.

Plus formellement, l'objectif est d'apprendre un modèle, ou une fonction, qui mappe un vecteur d'intrants à un vecteur d'extrants, compte tenu d'un ensemble d'exemples de formation (ou d'instances de formation) qui associe un vecteur d'intrants à ses extrants souhaités.

Cet ensemble de formation constitue l'expérience de formation. Enfin, la performance est mesurée par une fonction de perte, liée à l'erreur commise par le modèle lors de la prédiction de la valeur de sortie des exemples de test, différente des exemples de formation.

3.3.1 La Prédiction

Presque tous les algorithmes d'apprentissage automatique ont une chose en commun : ils sont alimentés avec des données et produisent une réponse à une question. Dans les cas extrêmes, les ensembles de données peuvent se composer de milliards de valeurs

alors que la réponse n'est qu'un seul bit. Par exemple, lorsqu'il faut prévoir les valeurs futures d'une série chronologique, il s'agit d'une prévision, et lorsque le contenu des images est reconnu, il s'agit d'une classification. Dans le contexte de la recherche médicale ou des applications l'ordinateur produit un diagnostic et si un robot doit agir dans un environnement on peut parler d'une décision.

Comme le résultat de l'algorithme est toujours lié à une certaine incertitude, une autre expression courante est une supposition. Ici, nous utiliserons principalement le terme extrant ou prédiction d'un modèle, tandis que les données sont considérées comme l'intrant.[32]

3.3.2 Régression ou Classification

On distingue usuellement les catégories de variables (prédictives ou cibles) suivantes :

- Les variables quantitatives sont des variables numériques qui correspondent à des quantités.
- Les variables qualitatives définissent des catégories ou des classes. On distingue deux sous 2 catégories [33] :
 - Les variables nominales - Ce sont des variables qualitatives définies par des noms, comme des noms pays par exemple.
 - Les variables ordinales - Ce sont des variables qualitatives que l'on peut comparer entre elles comme des niveaux de risques par exemple.

Ce qui nous amène aux deux définitions suivantes :

- Un modèle de régression est un modèle de ML dont la variable cible est quantitative. La régression linéaire est un exemple.
- Un modèle de classification est un modèle de ML dont la variable cible est qualitative. La régression logistique est un exemple où la variable cible est binaire.

3.3.3 Les Méthodes d'Apprentissage Supervisé

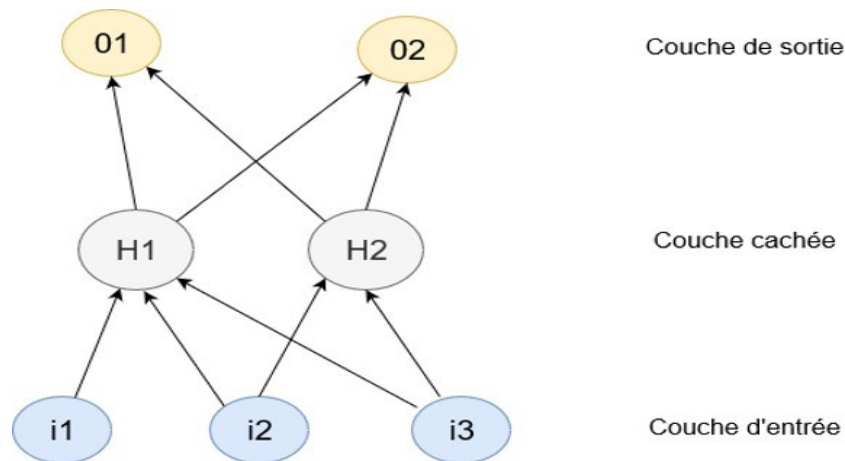
La Régression Logistique

La Régression logistique (RL) est une méthode de classification binaire à base statistique qui utilise un modèle linéaire pour effectuer une régression sur un ensemble de variables.

La RL est couramment utilisée pour la prédiction des tendances d'un ensemble de données avec des attributs numériques ou non ambigus.

Il utilise le logarithme pour calculer la probabilité à partir de plusieurs variables d'entrée.[34]

FIGURE 3.3 – Le modèle de régression logistique.



Source : Alkhateeb, Z. K. and Maolood, A. T. (2019). Machine learning-based detection of credit card fraud : A comparative study.[34]

Les Arbres de Décision

Les arbres de décision sont des techniques de ML qui expriment des attributs indépendants et un attribut dépendant dans une structure en forme d'arbre qui représente un ensemble de décisions.

Les règles de classification, extraites des arbres de décision, sont des expressions de type SI-ALORS dans lesquelles les conditions préalables sont logiquement ET et tous les tests doivent réussir pour que chaque règle soit générée.

L'arbre de décision utilise une combinaison d'arbres binaires et de nœuds lors de la clas-

sification des données.

Lorsque l'échantillon se déplace le long de l'arbre, les nœuds appartenant à cet échantillon sont générés. Ensuite, l'arbre est divisé en sous-ensembles et est ensuite stocké dans les "sous-groupes mutuellement exclusifs", c'est pourquoi il est appelé "arbre de classification et de régression".[35]

La Forêt aléatoire (Random Forest)

Les forêts aléatoires sont composées (comme le terme "forêt" l'indique) d'un ensemble d'arbres décisionnels binaires dans lequel a été introduit de l'aléatoire. Ces arbres se distinguent les uns des autres par le sous-échantillon de données sur lequel ils sont entraînés. Ces sous-échantillons sont tirés au hasard (d'où le terme "aléatoire") dans un jeu de données. La technique des forêts aléatoires modifie la méthode du Bagging appliquée ici aux arbres en ajoutant un critère de décorrélation entre ces arbres. L'idée de cette méthode est de réduire la corrélation sans augmenter trop la variance. Le principe consiste à choisir de façon aléatoire un sous-ensemble de variables qui sera considéré à chaque niveau de choix du meilleur nœud de l'arbre.[36]

Principe de La Forêt aléatoire

Considérons un ensemble d'entraînement $L = (X_1, Y_1), \dots, (X_m, Y_m)$, a le nombre d'attributs des exemples de X . Considérons également L_t un bootstrap contenant m instances obtenus par échantillonnage avec remplacement de L .

Soit h_1, \dots, h_T un ensemble de T arbres de décision. Chaque arbre h_t est construit à partir de L_t . Pour chaque nœud de l'arbre, l'attribut de partitionnement est choisi en considérant un nombre f ($f < a$) d'attributs choisis aléatoirement (parmi les a attributs). Pour classifier une nouvelle instance, le classificateur des forêts aléatoires effectue un vote de majorité uniformément pondéré des classificateurs de cet ensemble pour l'instance. L'algorithme illustre ce principe.

3.4 L'Apprentissage Non Supervisé

Ce serait une erreur de penser que l'apprentissage automatique exige toujours des exemples avec des étiquettes de classe. Loin de là ! Des informations utiles peuvent être glanées même à partir d'exemples dont les classes ne sont pas connues. C'est ce qu'on appelle parfois l'apprentissage non supervisé, contrairement au terme d'apprentissage supervisé qui est utilisé lorsqu'on parle d'initiation à partir d'exemples préclassifiés.

Alors que l'apprentissage supervisé se concentre sur l'induction des classificateurs, l'apprentissage non supervisé est intéressé à découvrir les propriétés utiles des données disponibles. Peut-être la tâche la plus populaire cherche des groupes (appelés clusters) d'exemples similaires. Les centroïdes de ces groupes peuvent alors être utilisés comme centres gaussiens pour les classificateurs bayésiens ou RBF, comme prédicteurs de valeurs d'attributs inconnus, et même comme outils de visualisation de données multidimensionnelles.

Enfin, les techniques utilisées dans l'apprentissage non supervisé peuvent être utilisées pour créer des attributs de niveau supérieur à partir des attributs existants.[37]

3.4.1 Clustering

Le clustering est une étude non supervisée, visant à organiser un ensemble d'objets en groupes ou clusters, de façon à avoir des objets similaires groupés et les objets différents organisés dans des groupes différents. Ce problème a été abordé dans de nombreux contextes et par des chercheurs dans beaucoup de disciplines, ce qui reflète son attrait et son utilité comme l'une des étapes les plus importantes de l'analyse exploratoire des données.

Le clustering est loin d'être trivial à réaliser. En fait, le problème est fondamentalement mal posé, c'est-à-dire un ensemble donné d'objets peuvent être regroupés d'une manière radicalement différente, sans avoir des critères pour préférer un regroupement plutôt qu'un autre.

En raison de l'ambiguïté intrinsèque concernant les problèmes du clustering, une vaste collection d'algorithmes s'est étendue dans la littérature afin d'améliorer les approches existantes sur des applications spécifiques.

Les techniques traditionnelles sont axées sur la notion de caractéristiques. Selon ce point de vue, chaque objet est décrit en termes d'un vecteur d'attributs numériques et est donc associé à un point dans un espace vectoriel (géométrique) Euclidien de sorte que les distances entre les points observées reflètent les dissimilarités/similarités entre les objets respectifs.

Cependant, l'inconvénient avec l'approche géométrique est sa limitation intrinsèque, qui porte sur le pouvoir de représentation de la caractéristique vectorielle, à base de descriptions. En fait, il existe de nombreux domaines d'application où soit il n'est pas possible de trouver les caractéristiques satisfaisantes où ils sont inefficaces pour des objectifs d'apprentissage.[38]

Le clustering est un processus qui regroupe un ensemble d'objets (physiques ou abstraits) en clusters similaires de telle sorte que les données du même cluster aient des caractéristiques similaires, et celles appartenant à des clusters distincts soient dissimilaires.[39]

3.4.2 Les Méthodes d'Apprentissage Non Supervisé

Les méthodes non supervisées sont utiles dans les applications où il n'y a pas de connaissance préalable quant à la classe particulière d'observations dans un ensemble de données. Par exemple, nous pouvons ne pas être en mesure de savoir avec certitude quelles transactions dans une base de données sont frauduleuses et lesquelles sont légitimes. Dans ces situations, des méthodes non supervisées peuvent être utilisées pour trouver des groupes ou des observations anormales dans les données.[40]

Essentiellement, nous collectons des données pour fournir un résumé du système que nous étudions. Une fois que nous disposons d'un résumé du comportement du système, nous pouvons identifier les observations qui ne correspondent pas à ce comportement, c'est-à-dire les observations anormales. C'est notre objectif en utilisant des techniques statistiques non supervisées pour la détection des fraudes. [41]

Les Méthodes Hiérarchiques

Dans un clustering hiérarchique, un cluster peut être divisé en sous clusters, l'ensemble des clusters étant généralement représentés par un arbre. Un objet appartient à une et une seule feuille dans la hiérarchie, mais également à son nœud père, et ainsi de suite jusqu'à la racine. Les méthodes de clustering hiérarchique permettent d'obtenir ce type de résultats.

Il existe deux types d'approches de clustering hiérarchique :

- Les approches par agglomération (ou ascendantes).
- Les approches par division (ou descendantes).

Les Méthodes de Partitionnement

Les méthodes de partitionnement ont généralement comme résultat un ensemble de M clusters, chaque objet appartenant à un seul cluster. Chaque cluster peut être représenté par un centroïde (représentant du cluster) qui peut être considéré comme une description récapitulative de tous les objets contenus dans le cluster. La forme précise de cette description dépendra du type des objets qui sont groupés.

Au cas où données à valeurs réelles seraient disponibles, la moyenne arithmétique des vecteurs d'attribut pour tous les objets dans un cluster fournit un représentant approprié, des types alternatifs de centroïdes peuvent être requis dans d'autres cas.

Si les nombres de clusters sont élevés, les centroïdes peuvent encore être groupés de manière hiérarchique.

Méthodes basées sur les grilles

Un algorithme de clustering basé sur les grilles utilise des structures de données Multirésolution, où l'espace d'objets est quantifié en un ensemble de cellules, puis identifie l'ensemble de cellules denses connectées pour former des clusters.

Méthodes basées sur la densité

Les algorithmes basés sur la densité sont capables de découvrir des clusters de formes arbitraires, ce qui assure l'isolement des bruits (outliers) et la prévention contre la

formation de clusters non pertinents.

Ces algorithmes regroupent des objets selon des fonctions de densité spécifiques. La densité est habituellement définie comme nombre d'objets dans un voisinage particulier des éléments de données. Dans cette approche, un cluster donné continue à augmenter de taille tant que le nombre d'objets dans le voisinage dépasse un certain seuil.

K_means

Le clustering K_means est un algorithme de clustering simple et largement utilisé. Avec une valeur de k , il tente de construire k grappes à partir d'échantillons de l'ensemble de données. Par conséquent, k est un hyperparamètre du modèle. La bonne valeur de k n'est pas facile à déterminer, car elle dépend fortement de l'ensemble de données et de la façon dont les données sont représentées.

Pour mesurer la similarité entre deux points de données, K_means nécessite la définition d'une fonction de distance entre les points de données. Qu'est-ce qu'une distance ? C'est une valeur qui indique la proximité de deux points de données dans leur espace.

En particulier, lorsque les points de données se trouvent dans un espace de dimension d , la distance euclidienne est un bon choix de fonction de distance.

En K_means, un cluster est un groupe de points, avec une entité représentative appelée un centroïde. Un centroïde est également un point dans l'espace de données : le centre de tous les points qui composent le cluster. Il est défini comme étant la moyenne arithmétique des points.

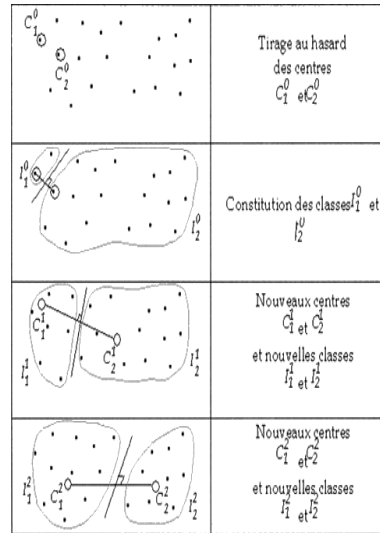
En général, lorsqu'on travaille avec des moyennes K, chaque échantillon de données est représenté dans un tableau numérique à d dimensions, pour lequel il est plus facile de définir une fonction de distance appropriée. Par conséquent, dans certaines applications, les données d'origine doivent être transformées en une représentation différente, pour répondre aux exigences des K_means.[41]

Étant donné k , l'algorithme de la moyenne K fonctionne comme suit :

1. Choisir au hasard k points de données (graines) comme étant les centroïdes initiaux
2. Assigner chaque point de données au centroïde le plus proche

3. Recalculer (mettre à jour) les centroïdes en utilisant les adhésions actuelles aux clusters
4. Si un critère de convergence n'est pas rempli, passez à l'étape 2

FIGURE 3.4 – Le schéma de K_means



Source : https://eric.univ-lyon2.fr/~ricco/cours/slides/classif_centres_mobiles. [42]

Nous pouvons également mettre fin à l'algorithme lorsqu'il atteint un budget d'itération, ce qui donne un résultat approximatif.

À partir du pseudo-code de l'algorithme, nous pouvons voir que les résultats de la mise en grappes de K_means peuvent être sensibles à l'ordre dans lequel les échantillons de données de l'ensemble de données sont explorés.

Une pratique sensée consisterait à exécuter l'analyse plusieurs fois, en randomisant l'ordre des objets ; ensuite, on fait la moyenne des centres de regroupement de ces passages et on entre les centres comme centres initiaux pour un passage final de l'analyse.

Les Avantages

L'algorithme du k_means converge en général très rapidement : il n'est pas rare qu'il atteigne la convergence au bout de 10 itérations, même avec beaucoup de points.

Comparée à l'utilisation d'autres méthodes de classification, une technique de classification k_means est rapide et efficace en termes de coût de calcul, en effet sa complexité est $O(K * n * d)$.

L'analyse par `K_means` améliore la précision de la classification et garantit que des informations sur un domaine de problème particulier sont disponibles. La modification de l'algorithme `k_means` basé sur ces informations améliore la précision des clusters.

Ce mode de regroupement fonctionne très bien lorsqu'il s'agit de clusters sphériques. Il fonctionne avec une hypothèse de distributions conjointes de caractéristiques puisque chaque cluster est sphérique. Toutes les caractéristiques ou tous les caractères des clusters ont la même variance et sont indépendants les uns des autres.[41]

Les Inconvénients

Malheureusement, le `k_means` n'est pas capable de déterminer le nombre de classes optimal : on est obligé de le lui spécifier au départ. Si on lui demande de trouver 3 clusters alors que vos données sont très clairement regroupées en 5 clusters, alors le `k_means` vous donnera 3 clusters, même si ce n'est visiblement pas la solution la meilleure.

Pour choisir le nombre optimal de clusters, on peut aussi lancer le `k_means` plusieurs fois, avec différentes valeurs de `K`. Pour chacune d'entre elles, on note l'inertie intraclasse obtenue. Bien entendu, l'inertie intraclasse diminue forcément quand `K` augmente, mais en observant la valeur de `K` au-delà de laquelle la diminution est plus faible, on peut déterminer une bonne valeur de `K`. C'est la méthode du coude.[41]

3.4.3 Choix du Nombre de Cluster

Définir le nombre de clusters est un des problèmes les plus difficiles en clustering. En effet, il est souvent nécessaire de fournir le nombre de clusters souhaité comme Paramètre de l'algorithme. Le choix du nombre de clusters a souvent été étudié comme un problème de sélection de modèle. Dans ce cas, l'algorithme est généralement exécuté plusieurs fois indépendamment avec un nombre de clusters différent. Les résultats sont ensuite comparés en se basant sur un critère de sélection qui permet de choisir la meilleure solution. Ce choix est toujours subjectif et fortement dépendant du critère sélectionné pour comparer les résultats.

Deux approches moins subjectives souvent utilisées se basent sur les critères de Minimum message Length (MML) et Minimum description Length (MDL). Elles consistent

à débiter avec un nombre de clusters relativement élevé, puis à fusionner itérativement deux clusters pour optimiser les critères (MML ou MDL). Les autres Critères classiquement utilisés pour la sélection de modèle sont le bayes information Criterion (BIC) et l'Akiake Information Criterion (AIC).

Le Gap statistics est également utilisé pour décider du nombre de clusters. Ces critères reposent généralement sur des bases statistiques fortes et s'appliquent de manière naturelle aux méthodes de clustering probabilistes. Elles peuvent être plus difficiles à mettre en place lors de l'utilisation d'autres types d'approches.

De plus, elles sont relativement couteuses et nécessitent d'effectuer de nombreuses exécutions des algorithmes. L'étude de la validité des clusters découverts peut également être un outil pour effectuer le choix du nombre de clusters. Dans l'idéal, le choix du nombre de clusters reste à l'appréciation de l'expert qui est à même, avec ou sans l'aide d'indices, de choisir le nombre de clusters qui lui paraît adapté.[43]

3.5 Prétraitement des Données

Le prétraitement des données est une étape importante dans tout projet d'exploration de données. Les données du monde réel sont généralement incomplètes, incohérentes et comportent même des erreurs. Afin de produire un bon modèle à un problème, ces questions doivent être abordées. Il sera discuté de plusieurs questions et techniques de prétraitement des données.

3.5.1 Problèmes de Données

Cette section classe les principaux problèmes de qualité des données à résoudre par le nettoyage et la transformation des données. Comme nous le verrons, ces problèmes sont étroitement liés et devraient donc être traités dans un moyen uniforme. Nettoyage d'un ensemble de données cible peut être une tâche encore plus lourde que la collection des données.[44] Les données peuvent contenir plusieurs types de problèmes :

- bruit : L'apparition de bruit dans les données est généralement attribuable à des erreurs d'enregistrement et à des limites technologiques. On peut également

établir un lien avec l'incertitude et la nature probabiliste des caractéristiques spécifiques et des valeurs de classe.

- **Données manquantes** : Des données manquantes peuvent survenir en raison de multiples situations. Il peut y avoir des conflits dans les données enregistrées, ce qui entraîne l'écrasement des données. Dans certains cas, ce champ particulier des données ne pouvait pas être considéré comme important à ce moment-là et n'a donc pas été saisi.
- **Données redondantes** : Ce type d'erreur est principalement lié à des erreurs humaines. Les données auraient pu être enregistrées sous différents noms ou à différents endroits. Elle peut également être représentative des enregistrements contenant des attributs non pertinents ou peu informatifs.
- **Données insuffisantes et périmées** : Parfois, les données dont nous avons besoin proviennent d'événements rares et, par conséquent, nous pouvons avoir des données insuffisantes. Parfois, les données peuvent ne pas être à jour et, par conséquent, nous pourrions devoir les jeter et nous retrouver avec des données insuffisantes.

3.5.2 Techniques de Prétraitement des Données

Afin d'améliorer la qualité des données recueillies précédemment, plusieurs techniques de prétraitement peuvent être appliquées.[44] Ces techniques peuvent être divisées en quatre grandes catégories : nettoyage des données, transformation des données, réduction des données et intégration des données.

Nettoyage des Données

Les techniques de nettoyage des données visent à nettoyer les données en remplissant les valeurs manquantes, en traitant les valeurs aberrantes, en lissant les données bruyantes et en corrigeant les incohérences. Lorsque les données contiennent des valeurs manquantes, quelques méthodes peuvent être mises en œuvre pour résoudre ce problème. Les approches les plus courantes pour résoudre ce problème sont d'ignorer le tuple, ou de remplir la valeur manquante en utilisant l'attribut moyen.

Si les données contiennent du bruit, certaines techniques de lissage des données peuvent

être appliquées, y compris les méthodes de regroupement et de régression. Les données incohérentes peuvent être corrigées au moyen d'une méthode de recherche sur papier. Les outils d'ingénierie des connaissances peuvent également être utilisés pour détecter les violations des contraintes de données connues.

Transformation des Données

La transformation des données consiste à reconstruire les données sous une nouvelle forme plus adaptée à la tâche d'exploration de données. Plusieurs types de transformation des données peuvent être utilisés :

- Normalisation
- Lissage
- Généralisation des données
- Agrégation

La normalisation est l'une des techniques les plus utilisées pour transformer les données. Dans le cas le plus simple, cela signifie ajuster les valeurs mesurées sur différentes échelles à une échelle commune. Dans de nombreux cas, cela se fait après la moyenne des chiffres.

Réduction des Données

L'exploration de grandes quantités de données peut être une tâche de longue date, parfois même irréalisable. Les techniques de réduction des données visent à réduire la quantité de données à analyser sans compromettre l'intégrité des données originale. Elle consiste à réduire le volume des données ou leur dimensionnalité, en supprimant les attributs. Il existe quelques stratégies de réduction des données :

- Agrégation de cubes de données
- Compression de données
- Réduction des données
- Réduction de numération
- Discrétisation

Intégration des Données

La plupart des projets d'analyse de données consistent à combiner des données provenant de sources multiples en un seul magasin de données.

Cette tâche s'appelle l'intégration des données. L'intégration de plusieurs sources de données présente plusieurs défis. Par exemple, le même attribut peut avoir des noms différents entre les différentes sources, et il peut ne pas avoir de nom intuitif.

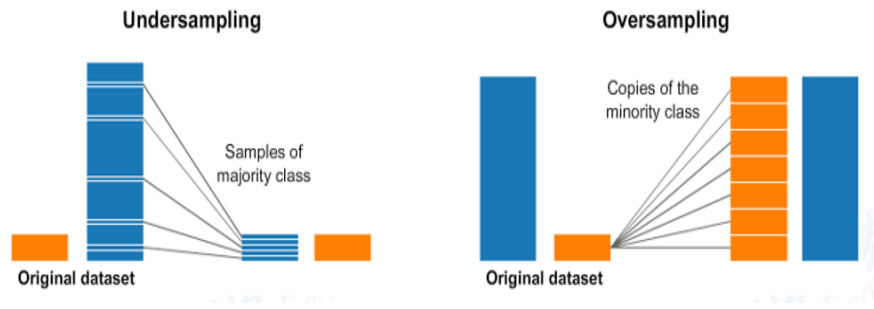
La meilleure approche dans ce genre de tâche est d'utiliser les métadonnées habituellement présentes dans les bases de données et les entrepôts pour aider à éviter les erreurs dans l'intégration.

3.6 Méthodes d'équilibrage des données

Les données déséquilibrées font référence aux types d'ensemble de données où la classe cible a une répartition inégale des observations, c'est-à-dire qu'une étiquette de classe a un très grand nombre d'observations et l'autre a un très faible nombre d'observations.

Supposons que XYZ est une banque qui émet une carte de crédit à ses clients. Maintenant, la banque est préoccupée par le fait que certaines transactions frauduleuses sont en cours et lorsque la banque vérifie ses données, elle constate que pour chaque transaction de 2000, il n'y a que 30 numéros de fraude enregistrés. Ainsi, le nombre de fraudes par 100 transactions est inférieur à 2 %, où nous pouvons dire que plus de 98% des transactions sont de nature « sans fraude ». Ici, la classe « Pas de fraude » est appelée la classe majoritaire, et la classe « fraude » beaucoup plus petite et appelée la classe minoritaire.[45]

FIGURE 3.5 – Méthodes d'équilibrage des données



Source : <https://www.pinterest.com/pin/514958538641697615/>. [46]

3.6.1 Méthodes de Over_Sampling

La méthode de suréchantillonnage fonctionne avec les cas de classe minoritaire qui ne nécessitent aucune perte d'information. Il est divisé en deux catégories :

Suréchantillonnage aléatoire

Échantillonnage InformatifOver.

Les méthodes de suréchantillonnage aléatoire stabilisent les ensembles de données avec l'échantillonnage aléatoire des instances de classe inférieure. Le suréchantillonnage informatif génère et ajoute les instances de classe minoritaires dans l'ensemble de données.

3.6.2 Méthodes de Under_Sampling

La méthode de sous-échantillonnage fonctionne avec les instances de classes majoritaires, qui suppriment les positions aléatoirement de la classe majoritaire pour équilibrer l'ensemble de données. Il est préférable de l'utiliser lorsque l'ensemble de données est important. Les méthodes de sous-échantillonnage sont de deux types :

Échantillonnage aléatoire.

Sous-échantillonnage informatif.

Le sous-échantillonnage aléatoire sélectionne de façon aléatoire les instances de la classe majoritaire, qui sont ensuite éliminées jusqu'à ce que l'ensemble de données soit équilibré. Le fait de supprimer des instances de façon aléatoire peut amener les données de formation à contenir des renseignements importants. Le sous-échantillonnage informatif utilise un facteur de sélection pré-spécifié pour supprimer les instances de classe

majoritaires.

3.7 Évaluation des modèles d'apprentissage automatique

Lorsque nous entraînerons notre algorithme d'apprentissage automatique sur notre ensemble de données et utiliserons les prédictions de ce même ensemble de données pour évaluer les algorithmes d'apprentissage automatique, nous serons confrontés à un problème général appelé le dépassement (overfitting).

Imaginez un algorithme qui se souvient de chaque observation qu'il montre pendant l'entraînement. Si nous évaluons notre algorithme d'apprentissage automatique sur le même ensemble de données utilisé pour former l'algorithme, alors un algorithme comme celui-ci aurait un score parfait sur l'ensemble de données de formation.

Mais les prédictions qu'il a faites sur de nouvelles données seraient terribles. Nous devons évaluer nos algorithmes d'apprentissage automatique sur des données qui ne sont pas utilisées pour former l'algorithme.

L'évaluation est une estimation que nous pouvons utiliser pour parler de la façon dont nous pensons que l'algorithme peut réellement faire dans la pratique. Ce n'est pas une garantie d'exécution. Une fois que nous avons estimé les performances de notre algorithme, nous pouvons ensuite ré-entraîner l'algorithme final sur l'ensemble des données de formation et le préparer pour une utilisation opérationnelle.[47] Ensuite, nous allons examiner la technique "Train and Test Sets" que nous pouvons utiliser pour diviser notre ensemble de données de formation et créer des estimations de performance utiles pour nos algorithmes d'apprentissage automatique.

3.7.1 Train_Test Split

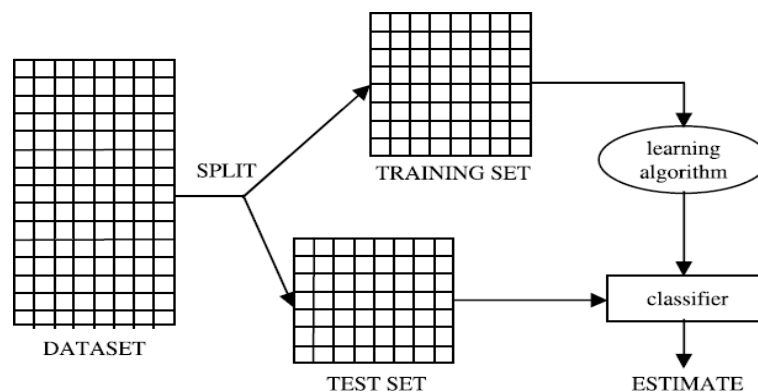
La procédure `train_test split` est utilisée pour estimer la performance des algorithmes d'apprentissage automatique lorsqu'ils sont utilisés pour faire des prédictions sur des données non utilisées pour former le modèle.

C'est une procédure rapide et facile à exécuter, dont les résultats vous permettent de comparer les performances des algorithmes d'apprentissage automatique pour votre problème de modélisation prédictive. Bien qu'elle soit simple à utiliser et à interpréter, il arrive que la procédure ne soit pas utilisée, par exemple lorsqu'il s'agit d'un petit ensemble de données et dans des situations où une configuration supplémentaire est requise, par exemple lorsqu'elle est utilisée pour la classification et que l'ensemble de données n'est pas équilibré.

Former l'algorithme sur la première partie, faire des prédictions sur la deuxième partie et évaluer les prédictions par rapport aux résultats attendus. La taille de la division peut dépendre de la taille et des spécificités de votre ensemble de données, dans notre cas, nous avons utilisé 60% des données pour la formation et les 40 % restants pour les tests.

Cette technique d'évaluation d'algorithme est très rapide. Il est idéal pour les grands ensembles de données (des millions d'enregistrements) où il existe des preuves solides que les deux fractionnements des données sont représentatifs du problème sous-jacent. En raison de la vitesse, il est utile d'utiliser cette approche lorsque l'algorithme que vous étudiez est lent à former.[47] La figure 3.6 montre les opérations de formation et de fractionnement.

FIGURE 3.6 – Les opérations de Train_Test Split



3.8 Méthodes d'ensemble

Chaque algorithme possède ses forces et ses faiblesses pour classifier de nouveaux exemples. En outre, certaines de ces forces sont complémentaires et peuvent améliorer les résultats lorsqu'elles sont combinées. Cette combinaison est surtout utile lorsque les

Classificateurs se trompent sur des exemples différents. Si les algorithmes font les mêmes erreurs, cette technique perd tout son lustre et ne change presque rien aux résultats. On peut citer les intérêts de la combinaison de Classificateur comme suivant :

- Distribuer les caractéristiques sur des Classificateurs adaptés.
- Exploiter la complémentarité entre Classificateur.
- Prendre en compte les performances de chacun des Classificateurs.
- Réduire l'importance des choix initiaux.
- Diviser pour mieux régner.

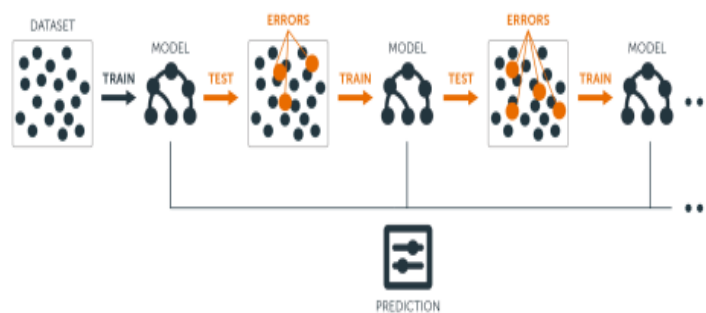
La combinaison se fait en entraînant séparément n modèles de manière à ce qu'ils soient les plus indépendants possibles l'un de l'autre et en combinant leurs réponses. Les méthodes de combinaison les plus communes sont le Bagging, et le Boosting.

3.8.1 La méthode de stimulation (Boosting)

La méthode du boosting est basé sur la création d'une famille de modèles qui sont ensuite agrégés par une moyenne pondérée des estimations. Elle diffère au niveau de la construction des modèles car chaque modèle est une version adaptative du précédent en donnant cette technique permet de produire des Classificateur "forts" (très précis) en combinant des instances "faibles" d'un Classificateur donné.[48]

L'algorithme AdaBoost, est le plus utilisé pour implémenter le boosting. Il construit de de façon itérative un ensemble de N Classificateur complémentaires.

FIGURE 3.7 – Schéma de la méthode de stimulation (Boosting)



Source :<https://machine-learning.paperspace.com/wiki/gradient-boosting>. [49]

Des Classificateurs faibles sont ajoutés si nécessaire, et entraînés sur les échan-

tillons que les Classificateurs précédents n'ont pas correctement classés. Les Classificateurs résultants sont combinés par un vote pondéré.

3.8.2 Classificateur de vote (Voting Classifier)

L'idée derrière le classificateur de vote est de combiner conceptuellement différents classificateurs d'apprentissage automatique et d'utiliser un vote à la majorité (Majority/-Hard voting) ou les probabilités prédites moyennes (soft voting) pour prédire les étiquettes de classe. Un tel classificateur peut être utile pour un ensemble de modèles tout aussi performants afin de compenser leurs faiblesses individuelles.[50]

Dans le vote à la majorité, l'étiquette de classe prévue pour un échantillon particulier est l'étiquette de classe qui représente la majorité (mode) des étiquettes de classe prédites par chaque classificateur individuel.

Par exemple, si la prédiction pour un échantillon donné est

Classificateur 1 -> classe 1

Classificateur 2 -> classe 1

Classificateur 3 -> classe 2

Voting Classifier (avec vote = 'Hard') classerait l'échantillon en tant que «classe 1» sur la base du libellé de la classe majoritaire.

En cas d'égalité, Voting Classifier sélectionnera la classe en fonction de l'ordre de tri croissant. Par exemple, dans le scénario suivant

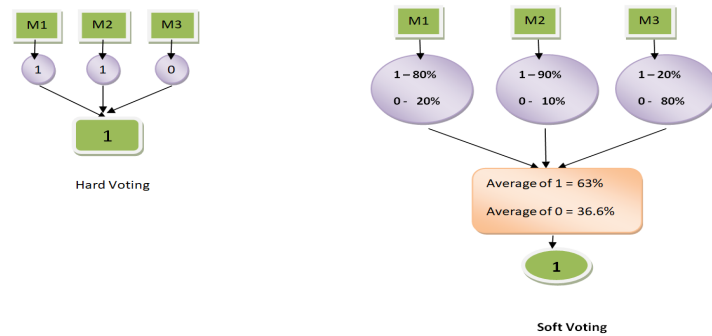
Classificateur 1 -> classe 2

Classificateur 2 -> classe 1

L'étiquette de classe 1 sera attribuée à l'échantillon. Contrairement au vote à la majorité (Hard Voting), le Soft Voting renvoie l'étiquette de classe en tant qu'argmax de la somme des probabilités prédites.

Des poids spécifiques peuvent être attribués à chaque classificateur via le paramètre poids. Lorsque des poids sont fournis, les probabilités de classe prédites pour chaque classificateur sont collectées, multipliées par le poids du classificateur et moyennées. L'étiquette de classe finale est ensuite dérivée de l'étiquette de classe avec la probabilité moyenne la plus élevée.

FIGURE 3.8 – Schéma de Classificateur de vote (Voting)



Source :<https://towardsdatascience.com/types-of-ensemble-methods-in-machine-learning-4ddaf73879db>. [51]

3.8.3 La méthode d'ensachage (Bagging)

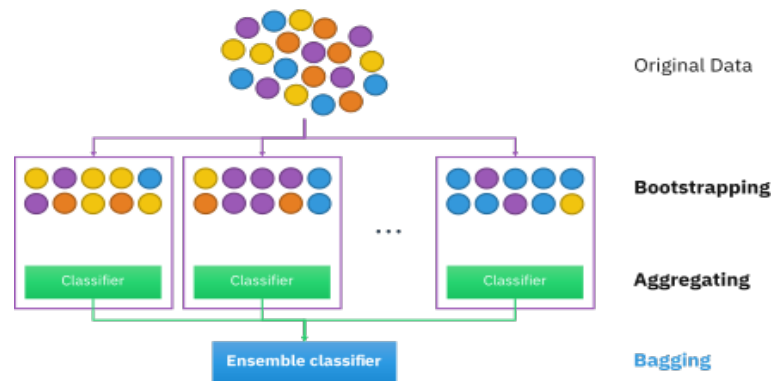
Bagging est une méthode proposée par pour construire des ensembles de Classificateur, chaque classificateur étant entraîné sur une réplique différente de la base d'apprentissage.[52] Les différentes répliques de la base d'apprentissage sont obtenues par bootstrap :

- Échantillon bootstrap : $l_1 = (x_1 \dots, x_n)$ est un échantillon aléatoire de taille n obtenu par tirage aléatoire avec remise dans l'échantillon original $l = (x_1 \dots, x_n)$.
- Chaque échantillon dans L peut apparaitre dans L^* zéro, une, deux, trois...fois.

Les différentes réplique l_i^* de la base d'apprentissage L sont très peu différentes de L mais suffisamment diverses pour obtenir des Classificateurs différents qu'on va pouvoir combiner.

La règle de combinaison peut être n'importe quelle règle. Par exemples on peut mettre la moyenne des prédictions des différents modèles.

FIGURE 3.9 – Schéma de la methode d'ensachage (Bagging)



Source : https://en.wikipedia.org/wiki/Bootstrap_aggregating. [53]

3.9 Les Métriques d'évaluation

Afin de pouvoir évaluer les performances d'un modèle de classification dans le monde de ML, nous nous intéressons aux indicateurs qui permettent de mesurer la qualité d'un modèle.

En général, pour évaluer la performance d'une solution, on divise l'échantillon de données disponibles en deux ensembles : l'ensemble d'apprentissage sur lequel le modèle fait son apprentissage et l'ensemble de tests sur lequel on peut évaluer sa performance. [54]

3.9.1 Matrice de confusion

La qualité d'un système de classification est mesurée à l'aide de la matrice de confusion.

Les colonnes de cette matrice représentent la répartition des objets dans les classes réelles. Les lignes quant à elles, représentent la répartition des points dans les classes estimées par un algorithme de classification. [54]

- Vrai positif VP : représente le nombre de résultat exprimant une opinion positive et classés positifs par le classifieur.
- Vrai négatif VN : représente le nombre de résultat exprimant une opinion négative et classés négatifs par le classifieur.
- Faux positif FP : représente le nombre de résultat exprimant une opinion

négative et classés positifs par le classifieur.

- Faux négatif FN : représente le nombre de résultat exprimant une opinion positive et classés négatifs par le classifieur.

FIGURE 3.10 – Matrice de Confusion

		Actual Value (as confirmed by experiment)	
		positives	negatives
Predicted Value (predicted by the test)	positives	TP True Positive	FP False Positive
	negatives	FN False Negative	TN True Negative

Source :

<https://medium.com/@natratanonkanraweekultana/confusion-matrix-d6146b275faa>. [55]

3.9.2 Précision

La probabilité conditionnelle qu'un exemple choisi aléatoirement soit bien classé par le système. Il s'agit du rapport entre le nombre de bonnes prédictions positives et le nombre de prédictions fausses positives. [55]

$$Precision = \frac{VP}{VP + FP} \quad (3.1)$$

3.9.3 Rappel

Mesure la largeur de l'apprentissage et correspond au rapport entre le nombre de bonnes prédictions positives et le nombre total d'exemples. [55]

$$Rappel = \frac{VP}{VP + FN} \quad (3.2)$$

3.9.4 F_mesure

La F-mesure est définie comme la moyenne harmonique de la précision et du rappel. Il s'agit d'une mesure qui est un compromis entre la précision et le rappel. [55]

$$F_{\text{mesure}} = \frac{2 * \textit{Precision} * \textit{Rappel}}{\textit{Precision} + \textit{Rappel}} \quad (3.3)$$

3.9.5 Taux de succès (exactitude) et taux d'erreur

Le taux de succès (A) et le taux d'erreur (B) sont deux mesures souvent utilisées par la communauté de ML. Le taux de succès (traduction de accuracy rate) désigne le pourcentage d'exemples bien classés par le classificateur, tandis que le taux d'erreur (error rate) désigne le pourcentage d'exemples mal classés.[55] Les deux taux sont estimés comme suit :

$$A = \frac{VP + VN}{VP + VN + FP + FN} \quad (3.4)$$

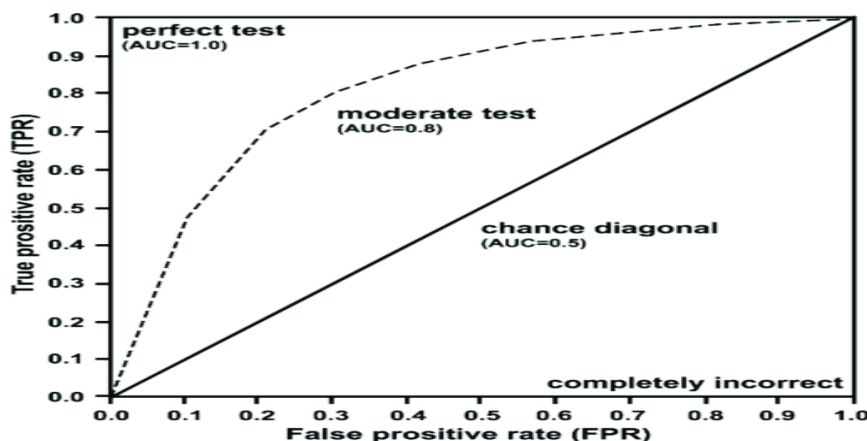
$$B = \frac{FP + FN}{VP + VN + FP + FN} = 1 - A \quad (3.5)$$

3.9.6 La courbe des caractéristiques d'exploitation du récepteur (Receiver Operating Characteristic Curve ROC)

La courbe des caractéristiques d'exploitation du récepteur (ROC) est une représentation des relations entre le taux positif réel (c.-à-d. les avantages) et le taux positif faux (c.-à-d. les coûts), tirés sur les axes x et y dans l'échelle plus mince. ROC représente les relations entre le rapport des churners correctement prédit comme churners, et le rapport des non-churners mal prédit comme churners. Le ROC fournit des compromis relatifs entre bénéfices et coûts. La courbe ROC se compose de points correspondant aux résultats de prédiction. La figure 3.11 présente un exemple de courbe ROC.

Le meilleur modèle de performance est lorsque la courbe ROC passe à travers ou près de (0, 1). La sensibilité et la spécificité du modèle seront alors de 100 % (c.-à-d. pas de faux négatifs et pas de faux positifs respectivement). Certains modèles comme la régression logistique produisent un score plutôt que de produire des décisions de classe binaire (c.-à-d. churn ou non-churn). Pour produire un classificateur binaire dans ce cas, des seuils sont utilisés. Si le résultat du classificateur est supérieur à un seuil, la classe de classification est un churn. Autrement, elle n'est pas un churn. La figure 3.11 illustre

FIGURE 3.11 – La courbe des caractéristiques d’exploitation du récepteur



Source : https://www.researchgate.net/figure/Receiver-operating-characteristic-ROC-curves-with-respective-area-under-the-curve-AUC_fig3_331311273. [56]

la courbe ROC pour un prédicteur aléatoire représenté par la ligne diagonale qui divise l’espace ROC en deux parties représentées. Les courbes ROC passant près de cette ligne correspondent à des classificateurs aléatoires (p. ex., classement par tirage à pile ou face).

En général, les modèles dont les courbes ROC passent la partie supérieure gauche de la courbe ROC obtiennent de meilleurs résultats. La zone sous la courbe ROC (appelée AUC) est utilisée comme mesure de performance. La valeur AUC varie de 0,0 à 1,0. Les modèles obtiennent de meilleurs résultats lorsque l’AUC est plus élevée. De plus, les modèles dont la valeur AUC est supérieure à 0,5 donnent de meilleurs résultats que les modèles aléatoires parce que la zone sous la ligne diagonale ROC est de 0,5. [56]

3.10 Conclusion

Le travail réalisé dans ce chapitre nous a permis de rassembler les différents outils techniques nécessaires à la réalisation de notre projet de fin d’études. Nous avons ainsi pu nous familiariser avec ce qui a été accompli dans le domaine de la prédiction de churn et les différents aspects de ses solutions.

En effet, en faisant une étude sur les méthodes que les analystes de données ont mises au point pour la détection de churn, nous pouvons distinguer les solutions les mieux adaptées à notre cas, ce qui nous permettra de les combiner et même de les développer

davantage au cas où nous le jugerions nécessaire.

Chapitre 4

Conception

4.1 Introduction

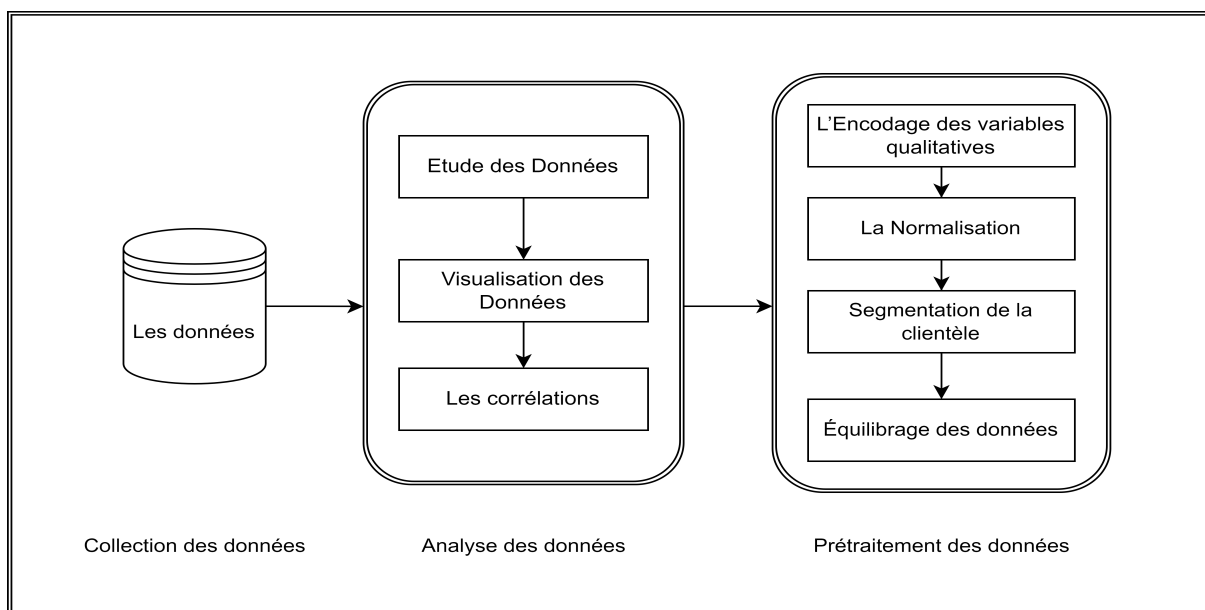
Les modèles qui prédisent le taux de désabonnement des clients sont basés sur la connaissance des clients de l'entreprise et de leurs appels. Ces informations sont stockées dans un fichier CSV. Tous les modèles ont été formés et testés avec ces données. Mais avant que cela puisse arriver, cette collection a dû passer par de multiples transformations et des techniques de prétraitement pour la prédire.

Ce chapitre vise à décrire toutes les étapes concernant les données. Il commence à expliquer comment les données ont été récupérées et stockées, comment nous avons mené le processus de sélection de l'information nécessaire, son analyse et les principales transformations effectuées pour la prédire.

4.2 Architecture du système

pour l'architecture on va suivre de l'organigramme présenté par la figure 4.1 :

FIGURE 4.1 – L'organigramme d'architecture du système



4.2.1 Collection des Données

Les données examinées dans notre étude font partie d'une compétition (Expresso churn Prediction challenge) sont recueillis et stockés dans un fichier CSV appelé "Train".

Le jeu de données final généré contient les opérations téléphoniques effectuées par plus de 2 millions de clients. Il existe une colonne appelée CHURN qui indique si un client a quitté l'entreprise ou non.

Afin de réaliser notre étude, nous nous sommes disposés d'un échantillon composé de 2154048 clients répartis comme suit :

- 403986 des clients churn.
- 1750062 des bons clients.

4.2.2 Analyse des Données

Après avoir récupéré les données, nous avons d'abord procédé à une visualisation globale de celles-ci pour avoir une idée de ce que nous avons à notre disposition comme matière première, et identifier les problèmes et les incohérences présentes.

Etude des Données

Nous avons commencé par citer les colonnes et leurs types, le nombre de lignes, et nous avons également fait des statistiques basiques (min, max, moyenne, ..etc) pour chaque colonnes, visualisation des premières et des dernières lignes par les fonctions `head()` et `tail()`, visualisation de chaque colonne a part, ...etc.

Cette étape est très importante car elle représente la première vue de ce que nous avons et elle nous permet d'avoir une idée générale et de mieux connaître nos données et leurs types.

L'ensemble de données disponible pour cette recherche contient des informations concernant les appels des clients.

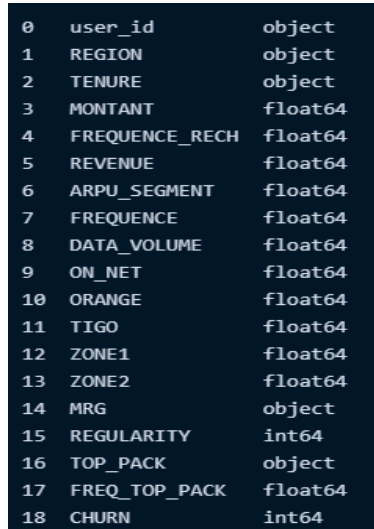
Le tableau 4.1 contient un aperçu des données. Il identifie les variables, leur variable et une simple description de leur signification.

VARIABLES	VALEUR	DESCRIPTION
REGION	Qualitative	la localit� de chaque client
TENURE	Quantitative	la dur�e dans le r�seau
MONTANT	Quantitative	montant de recharge
FREQUENCE-RECH	Quantitative	nombre de fois que le client a fait une recharge
REVENUE	Quantitative	revenu mensuel de chaque client
ARPU-SEGMENT	Quantitative	revenu sur 90 jours/3
FREQUENCE	Quantitative	nombre de fois que client � fait un revenu
DATA-VOLUME	Quantitative	nombre de connexions
ON-NET	Quantitative	nombre des appels inter expresso
ORANGE	Quantitative	nombre des appels vers orange
TIGO	Quantitative	nombre des appels vers Tigo
ZONE1	Quantitative	nombre des appels vers les zone1
ZONE2	Quantitative	nombre des appels vers les zone2
MRG	Qualitative	un client qui fait du vas
REGULARITY	Quantitative	nombre de fois que le client est actif pendant 90 jours
TOP-PACK	Qualitative	les pack les plus activ�s
FREQ-TOP-PACK	Quantitative	nombre de fois que le client a activ� les packages top pack
CHURN	0 ou 1	0 si le client est rest�, 1 si le client est parti

TABLE 4.1 – Les variables du jeu de donn es.

Afin d'explorer les données et aussi pour implémenter et évaluer les algorithmes, nous avons choisi le langage Python. Ces données ont été stockées dans un bloc de données et contiennent des valeurs catégorielles (caractère) et continues (numériques). Une variété de fonctions utiles pour explorer la trame de données ont été appliquées, à savoir les fonctions `info()` et `describe()`.

FIGURE 4.2 – Structure de Données



```
0  user_id      object
1  REGION      object
2  TENURE      object
3  MONTANT     float64
4  FREQUENCE_RECH float64
5  REVENUE     float64
6  ARPU_SEGMENT float64
7  FREQUENCE   float64
8  DATA_VOLUME float64
9  ON_NET      float64
10 ORANGE      float64
11 TIGO        float64
12 ZONE1       float64
13 ZONE2       float64
14 MRG         object
15 REGULARITY  int64
16 TOP_PACK    object
17 FREQ_TOP_PACK float64
18 CHURN       int64
```

La fonction `info()` renvoie un affichage compact de la structure interne des données. Il renvoie quelques exemples de sorties et les types de données pour chaque colonne. Comme représente la figure 4.2, il existe trois types de variables : (`object`) objet, (`float`) réelles (numérique), et (`category`) catégorie, qui sont les variables entières, et `factor`, représentant les variables catégorielles.

FIGURE 4.3 – Description des Données

	MONTANT	FREQUENCE_RECH	REVENUE	ARPU_SEGMENT	FREQUENCE	DATA_VOLUME	ON_NET
count	1.397309e+06	1.397309e+06	1.428000e+06	1.428000e+06	1.428000e+06	1.093615e+06	1.367373e+06
mean	5.532117e+03	1.152912e+01	5.510810e+03	1.836943e+03	1.397814e+01	3.366450e+03	2.776891e+02
std	7.111339e+03	1.327407e+01	7.187113e+03	2.395700e+03	1.469403e+01	1.330446e+04	8.726889e+02
min	1.000000e+01	1.000000e+00	1.000000e+00	0.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00
25%	1.000000e+03	2.000000e+00	1.000000e+03	3.330000e+02	3.000000e+00	0.000000e+00	5.000000e+00
50%	3.000000e+03	7.000000e+00	3.000000e+03	1.000000e+03	9.000000e+00	2.570000e+02	2.700000e+01
75%	7.350000e+03	1.600000e+01	7.368000e+03	2.456000e+03	2.000000e+01	2.895000e+03	1.560000e+02
max	4.700000e+05	1.330000e+02	5.321770e+05	1.773920e+05	9.100000e+01	1.823866e+06	5.080900e+04

	ORANGE	TIGO	ZONE1	ZONE2	REGULARITY	FREQ_TOP_PACK	CHURN
count	1.258800e+06	864032.000000	169721.000000	136824.000000	2.154048e+06	1.251454e+06	2.154048e+06
mean	9.541871e+01	23.109253	8.170132	7.553309	2.804251e+01	9.272461e+00	1.875474e-01
std	2.049873e+02	63.578086	41.169511	33.487234	2.228686e+01	1.228044e+01	3.903504e-01
min	0.000000e+00	0.000000	0.000000	0.000000	1.000000e+00	1.000000e+00	0.000000e+00
25%	7.000000e+00	2.000000	0.000000	0.000000	6.000000e+00	2.000000e+00	0.000000e+00
50%	2.900000e+01	6.000000	1.000000	2.000000	2.400000e+01	5.000000e+00	0.000000e+00
75%	9.900000e+01	20.000000	3.000000	5.000000	5.100000e+01	1.200000e+01	0.000000e+00
max	2.132300e+04	4174.000000	4792.000000	3697.000000	6.200000e+01	7.130000e+02	1.000000e+00

La fonction `describe()` est une fonction générique utilisée pour produire des résumés de résultats de données. Lorsqu'il est appliqué à une trame de données, il est appliqué à chaque colonne et les résultats de toutes les colonnes sont affichés ensemble.

La fonction `describe()` du framework Pandas énumère 8 propriétés statistiques de chaque attribut. Elles sont :

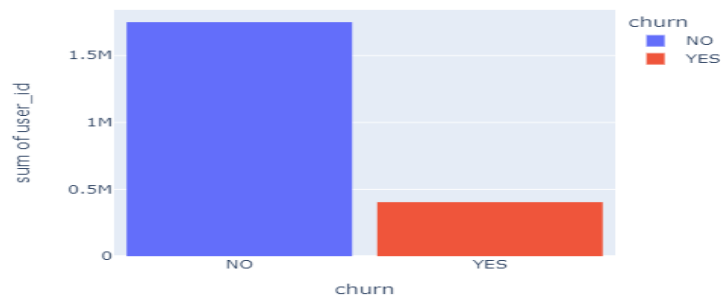
- . Compter.
- . Moyenne.
- . Écart-type.
- . Valeur minimale.
- . 25e percentile.
- . 50e centile (médiane).
- . 75e centile.
- . Valeur maximale.

Visualisation des Données

Distribution de Variable Churn

La répartition finale des clients concernant le taux de churn est représentée dans la figure 4.4.

FIGURE 4.4 – La Distribution de la variable churn



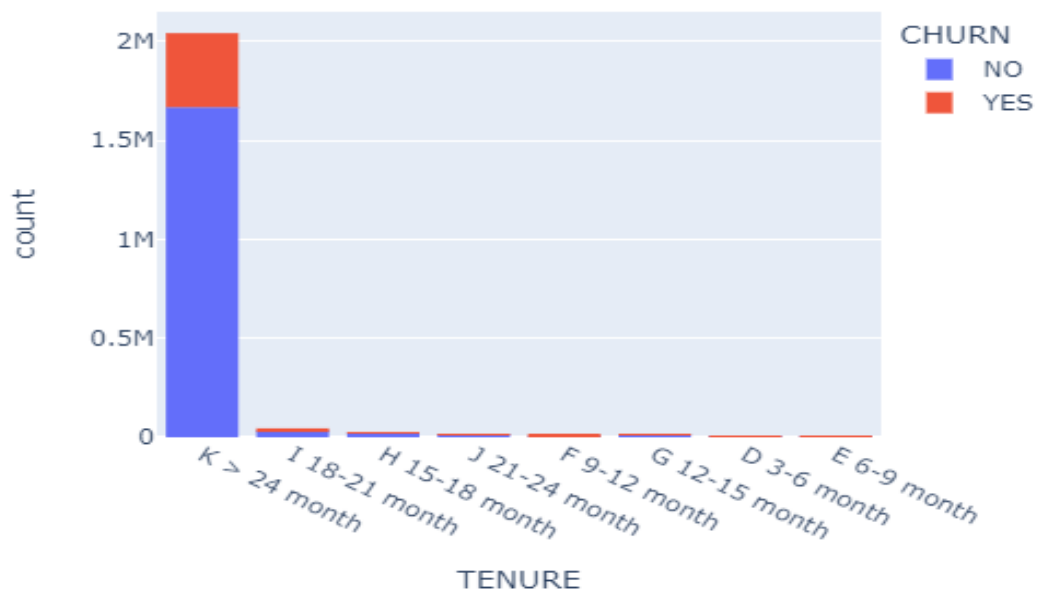
Ceci est notre variable cible, ce qui signifie que l'objectif des modèles développés est de prédire le résultat de cette variable à travers l'étude des autres variables de nos données. Nous pouvons déjà affirmer que notre classe cible est déséquilibrée, ce qui signifie que les classes ne sont pas représentées de manière égale.

Avec ces résultats, nous pouvons également voir le taux de désabonnement réel (observé) et qu'en divisant le nombre de churners par le nombre total de non-churners $\text{Churners} / \text{Non-churners} + \text{churners}$ Comme nous pouvons le voir à partir de ces résultats 18.8% de tous les clients sont des churners, tandis que le reste de la proportion va aux non-churners.

Répartition du churn selon l'ancienneté des abonnés

La répartition des clients churn selon l'ancienneté des abonnés, représenté par la figure 4.5 :

FIGURE 4.5 – Répartition du churn selon l'ancienneté des abonnés



Puisque la majorité des clients de ce jeu de données ont d'ancienneté de plus de deux ans alors le nombre des clients churners dans cette période est élevé. Mais pour les autres périodes il y a presque la même proportion des clients churn.

Les corrélations

Matrice des corrélations entre les fonctionnalités

La corrélation donne une indication de la façon dont les changements sont liés entre deux variables. Si deux variables changent dans la même direction, elles sont positivement corrélées. S'ils changent dans des directions opposées ensemble (on monte, on descend), alors ils sont négativement corrélés. La figure 4.6 montre une matrice de corrélation entre les caractéristiques numériques de notre ensemble de données.

FIGURE 4.6 – Matrice des corrélations entre les fonctionnalités

	MONTANT	FREQUENCE_RECH	REVENUE	ARPU_SEGMENT	FREQUENCE	DATA_VOLUME	ON_NET	ORANGE	TIGO	ZONE1	ZONE2	REGULARITY	FREQ_TOP_PACK
MONTANT	1.000000	0.793086	0.975393	0.975393	0.775892	0.295925	0.327328	0.658967	0.412771	0.366804	0.418742	0.522909	0.740771
FREQUENCE_RECH	0.793086	1.000000	0.801170	0.801169	0.956415	-0.149788	0.405065	0.519180	0.341972	0.119140	0.138243	0.557934	0.867009
REVENUE	0.975393	0.801170	1.000000	1.000000	0.786404	0.302907	0.332540	0.661516	0.412016	0.375026	0.375099	0.532654	0.750919
ARPU_SEGMENT	0.975393	0.801169	1.000000	1.000000	0.786403	0.302907	0.332540	0.661516	0.412016	0.375026	0.375099	0.532653	0.750919
FREQUENCE	0.775892	0.956415	0.786404	0.786403	1.000000	0.166126	0.395238	0.472231	0.306486	0.109066	0.147192	0.591588	0.843039
DATA_VOLUME	0.295925	0.149788	0.302907	0.302907	0.166126	1.000000	-0.013004	0.062921	0.022668	0.033459	0.051667	0.182272	0.114940
ON_NET	0.327328	0.405065	0.332540	0.332540	0.395238	-0.013004	1.000000	0.219474	0.136618	0.003501	-0.015299	0.270190	0.355803
ORANGE	0.658967	0.519180	0.661516	0.661516	0.472231	0.062921	0.219474	1.000000	0.403205	0.048939	0.034955	0.308198	0.553025
TIGO	0.412771	0.341972	0.412016	0.412016	0.306486	0.022668	0.136618	0.403205	1.000000	0.011690	0.017819	0.193812	0.367489
ZONE1	0.366804	0.119140	0.375026	0.375026	0.109066	0.033459	0.003501	0.048939	0.011690	1.000000	0.087099	0.045663	0.203852
ZONE2	0.418742	0.138243	0.375099	0.375099	0.147192	0.051667	-0.015299	0.034955	0.017819	0.087099	1.000000	0.063810	0.072849
REGULARITY	0.522909	0.557934	0.532654	0.532653	0.591588	0.182272	0.270190	0.308198	0.193812	0.045663	0.063810	1.000000	0.445550
FREQ_TOP_PACK	0.740771	0.867009	0.750919	0.750919	0.843039	0.114940	0.355803	0.553025	0.367489	0.203852	0.072849	0.445550	1.000000

Selon la matrice de corrélation entre les attributs, on observe une forte corrélation positive entre revenu et MONTANT, $\text{corr}(\text{REVENUE}, \text{MONTANT}) = 0.97$ et deux autres forte corrélation entre REVENUE et ARPU-SEGMENT, $\text{corr}(\text{REVENUE}, \text{ARPU-SEGMENT}) = 0.97$ puisque ses caractéristiques sont relativement homogènes dans chaque situation.

Les corrélations avec la variable cible

FIGURE 4.7 – Les corrélations avec la variable cible

```

REGULARITY      -0.479991
FREQUENCE       -0.139363
FREQUENCE_RECH  -0.123439
ARPU_SEGMENT    -0.114079
REVENUE         -0.114079
MONTANT         -0.105046
FREQ_TOP_PACK   -0.085106
ORANGE          -0.063400
ON_NET          -0.058698
TIGO            -0.035668
DATA_VOLUME     -0.032422
ZONE2           0.003379
ZONE1           0.009724
CHURN           1.000000
Name: CHURN, dtype: float64

```

Selon la figure 4.7, nous pouvons clairement voir qu'il y a une corrélation positive considérable entre la caractéristique ZONE2 et le churn, nous observons que la majorité des caractéristiques sont corrélées avec la variable cible négativement comme la RÉGULARITÉ et le churn avec un taux de corrélation de 0,48.

4.2.3 Le Prétraitement des données

Le prétraitement des données (Preprocessing) c'est l'étape qui consiste à préparer nos données avant de les fournir à la machine pour son apprentissage, le but a mis les données dans un format propice aux machines Learning afin d'améliorer la performance de ses modèles.

Parmi les opérations de prétraitement les plus importantes, que nous avons appliqué à l'ensemble de données :

L'Encodage des variables qualitatives

Avant d'aller plus loin, nous devons traiter des variables catégoriques. Un modèle d'apprentissage automatique ne peut malheureusement pas traiter les variables catégorielles (à l'exception de certains modèles tels que LightGBM mais il nécessite plus de 16 G RAM) par conséquent, nous devons trouver un moyen de coder (représenter) ces variables sous forme de nombres avant de les transférer au modèle.

Il y a deux manières principales de mener à bien ce processus :

- Label Encoder : attribue chaque catégorie unique dans une variable catégorielle avec un entier.
- One-hot encoding : crée une nouvelle colonne pour chaque catégorie unique dans une variable catégorielle.

Nous avons utilisé la politique suivante : pour toute variable catégorielle ayant plus de 2 catégories uniques nous utiliserons le codage d'étiquette (label encoding).

La Normalisation

Nous avons normalisé nos données afin que toutes les données puissent être égales en matière d'unité de mesure. Une façon simple d'y parvenir était d'utiliser une fonction

Scikit-learn intégrée appelée MinMaxScaler.

Normalisez les valeurs scalaires réelles des jeux de données avec les valeurs maximales et minimales à l'aide de l'équation :

$$new(v) = \frac{v - \min(e)}{\max(e) - \min(e)} \quad (4.1)$$

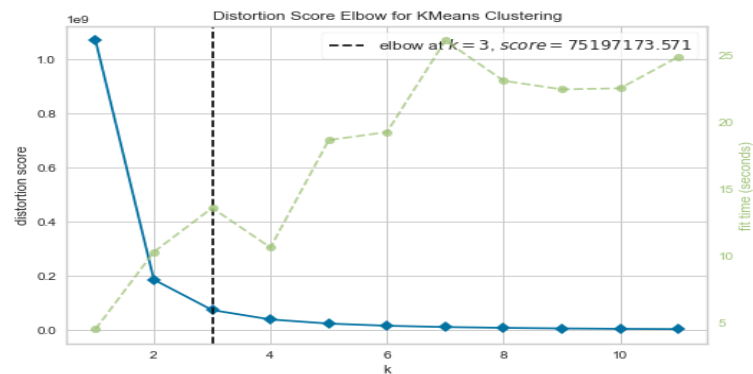
$\min(e)$ et $\max(e)$ sont les valeurs minimale et maximale de l'attribut E.

Segmentation de la clientèle

Pour appliquer l'algorithme `k_means`, nous avons choisi de grouper les clients selon la régularité où elle présente le nombre de fois que le client est actif pendant 90 jours. Et nous avons choisi cet attribut puisque la probabilité de désabonnement des clients qui deviennent inactifs et ne fasse aucune transaction pendant 90 jours, est très élevé.

Nous devons d'abord déterminer combien de groupes nous voulons. on va commencer par 4, mais pour être complet, nous allons utiliser l'inertie pour nous aider avec cette décision.

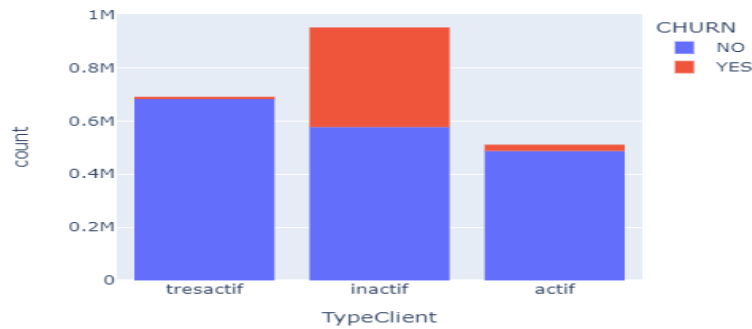
FIGURE 4.8 – Le diagramme d'inertie



Typiquement, nous regardons ce diagramme d'inertie pour trouver le point de coude "Elbow point". Dans notre cas, il semble que le coude se produit avec 3 groupes, alors nous allons procéder avec cela.

Et créons notre modèle final avec 3 catégories de client : très actif, actif, inactif. Comme le montre la figure 4.9, les clients inactifs sont plus susceptibles de se désabonner.

FIGURE 4.9 – Les types des clients



Équilibrage des données

Lorsqu'il s'agit d'un problème de classification, il se peut que nous n'obtenions pas toujours le ratio cible de façon égale. Il y aura des situations où vous obtiendrez des données très déséquilibrées, c'est-à-dire non égales. Dans le monde de l'apprentissage automatique, il est appelé cela un problème de données déséquilibrées de classe.

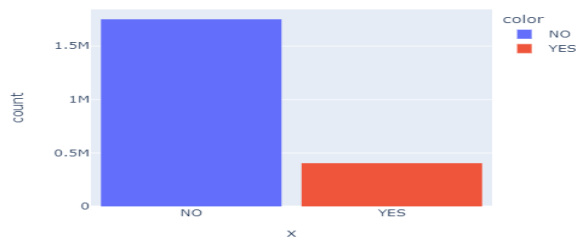
La construction de modèles pour les données cibles équilibrées est plus confortable que la manipulation de données déséquilibrées, même les algorithmes de classification trouvent plus facile d'apprendre de données correctement équilibrées.

Mais dans le monde réel, les données ne sont pas toujours fructueuses pour construire des modèles facilement. Nous devons gérer les données non structurées, et nous devons gérer les données de déséquilibre.

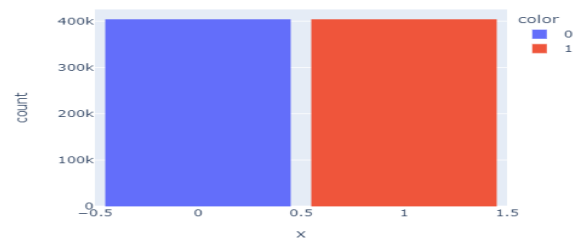
Et ça s'est-ce le cas dans notre étude les données sont clairement déséquilibrées (la figure 4.10).

Nous avons adapté comme une solution la méthode `Under_sampling` le type sur-échantillonnage aléatoire (`Random under sampling`).

FIGURE 4.10 – l'Équilibrage des données



((a)) Avant l'Équilibrage des données



((b)) Après l'Équilibrage des données

4.3 Conclusion

Dans ce chapitre, nous avons donné une description du jeu de données exploitées dans notre travail après nous avons exposé la répartition de la variable à prédire Churn et puis nous avons présenté les transformations appliquées sur l'ensemble de données. Après nous avons proposé un regroupement (clustering) des clients avec l'algorithme K_means .

Dans le chapitre suivant nous allons présenter les résultats des modèles proposés de manière plus détailler.

Chapitre 5

Réalisation et Tests

5.1 Introduction

dans ce chapitre nous présentons les résultats et les tests de notre solution ainsi que les techniques adéquates utilisés.

5.2 Évaluation des modèles

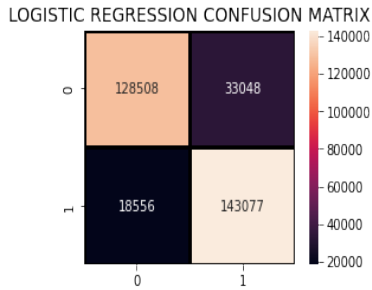
Pour notre phase de Évaluation des modèles, nous avons décidé de diviser nos données en deux parties, une partie de 60 % pour le processus de formation et 40 % pour la mise à l'essai et l'évaluation des prévisions par rapport aux résultats attendus. La méthode utilisée pour y parvenir se trouve dans le pack Scikit-learn appelé `train_test_split()`.

5.2.1 Matrice de confusion

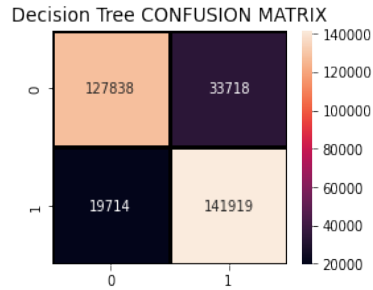
- TP (True Positive) : Le nombre de non-chuners que notre modèle a classés et prédit correctement.
- FN (False Negative) : Le nombre de chuners que notre modèle a classés comme des churners et prédit comme non-churners.
- FP (False Positive) : Le nombre de non-churners que notre modèle a redicte en tant que churners.

- TN (True Negative) : Le nombre de churners que notre modèle a correctement redicte comme churners.

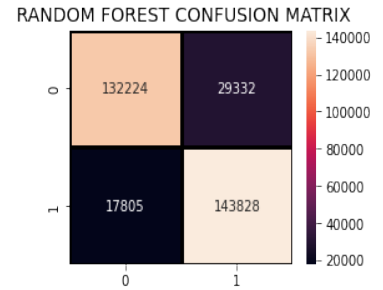
FIGURE 5.1 – Matrice de confusion



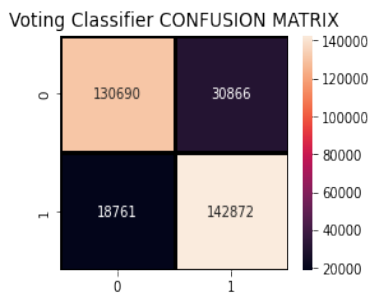
((a)) La Régression logistique



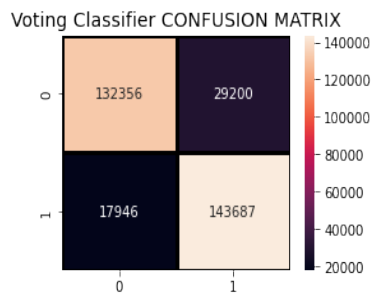
((b)) Classificateur de l'arbre décisionnel



((c)) La Forêt aléatoire



((d)) Classificateur de vote doux (soft)



((e)) Classificateur de vote difficile (hard)

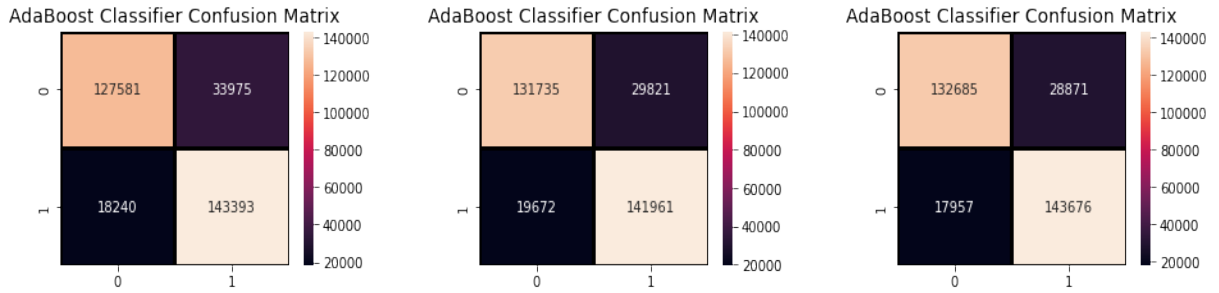
Les modèles	RL	DTC	RF	VCS	VCH
accuracy	0.840	0.834	0.854	0.846	0.854
Précision	0.812	0.808	0.830	0.822	0.831
Rappel	0.885	0.878	0.889	0.883	0.888
F-mesure	0.847	0.841	0.859	0.852	0.859
temps d'exécution	9s	7s	3m16s	3m6s	3m7s

TABLE 5.1 – Comparaison des modèles

Les meilleurs modèles selon le tableau sont : Forêt aléatoire, classificateurs de vote (les deux types).

La méthode de stimulation (boosting)

FIGURE 5.2 – Matrice de confusion de la méthode de stimulation



((a)) la stimulation de la Régression logistique

((b)) la stimulation de l'arbre décisionnel

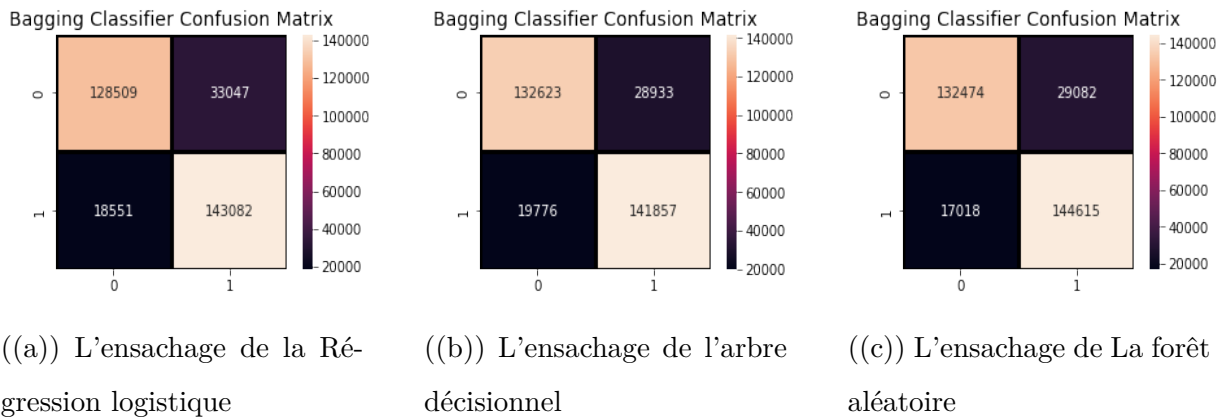
((c)) la stimulation de La forêt aléatoire

Le boosting de	RL	DTC	RF
accuracy	0.838	0.846	0.855
Précision	0.808	0.826	0.832
Rappel	0.887	0.878	0.888
F-mesure	0.845	0.851	0.859
temps d'exécution	1m	3m4s	47m39s

TABLE 5.2 – Comparaison des résultats de méthode de stimulation

La méthode d'ensachage (Bagging)

FIGURE 5.3 – Matrice de confusion de méthode d'ensachage



Le bagging de	RL	DTC	RF
accuracy	0.840	0.849	0.857
Précision	0.812	0.830	0.832
Rappel	0.885	0.877	0.894
F-mesure	0.847	0.853	0.862
temps d'exécution	1m	53s	18m11s

TABLE 5.3 – Comparaison des résultats de méthode d'ensachage

5.2.2 Courbe ROC

L'AUC est une métrique d'évaluation commune pour les problèmes de classification binaire. Il représente la valeur de zone créée par la courbe ROC. Si un classificateur est Très bon, le taux positif réel augmentera rapidement et l'aire sous la courbe sera proche de 1. Si le classificateur est similaire à une estimation aléatoire, le taux positif réel augmentera linéairement avec le taux de faux positifs et la zone sous la courbe sera d'environ 0,5.

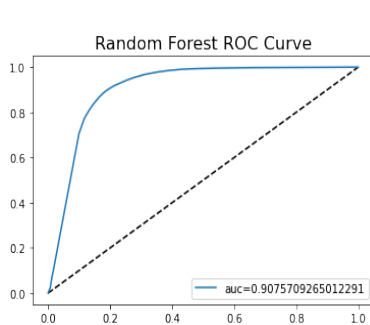
Les valeurs concernant AUC pour notre modèle sont formées avec une représentation visuelle des valeurs ROC de tous les modèles.

Mais si l'analyse s'arrête seulement sur la comparaison des courbes ROC, on remarque très bien que les performances des différents modèles sont similaires.

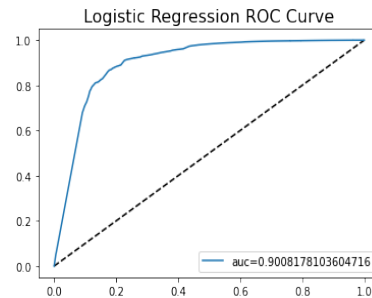
En effet, les mesures spécificité et sensibilité des différents modèles sont toutes significativement différentes de 0, nous pouvons donc en déduire que tous les modèles sont productibles et fiables.

Les modèles étudiés, deux d'entre eux ont atteint de bonnes valeurs AUC : Forêt aléatoire, et la régression logistique.

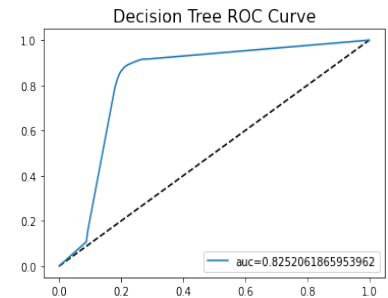
FIGURE 5.4 – Courbe ROC



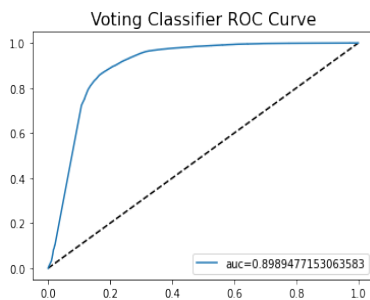
((a)) La Forêt aléatoire



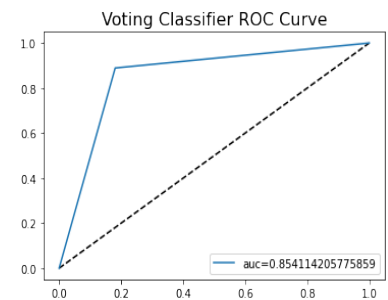
((b)) La Régression logistique



((c)) Classificateur de l'arbre décisionnel



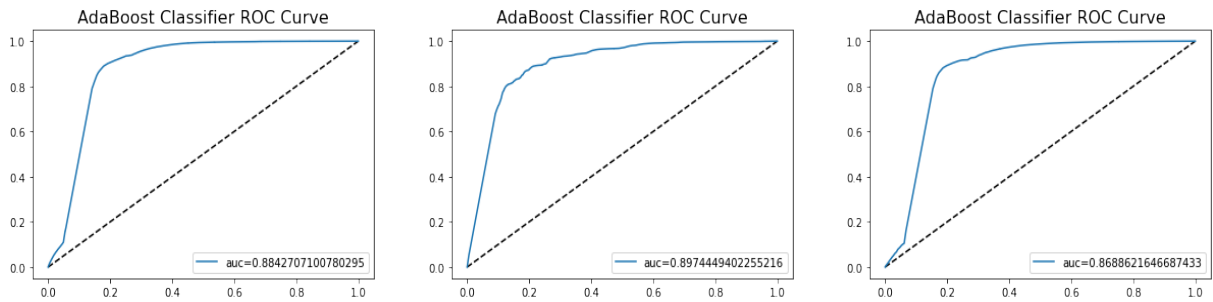
((d)) Classificateur de vot doux



((e)) Classificateur de vote difficile

La méthode de stimulation (boosting)

FIGURE 5.5 – Courbe ROC de la méthode de stimulation



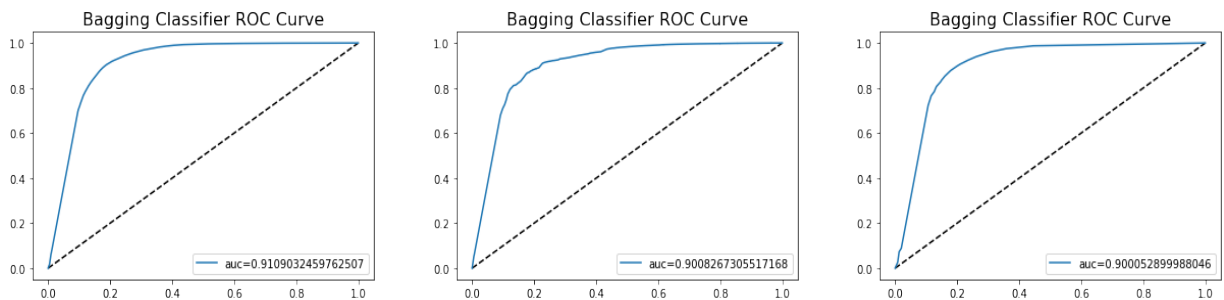
((a)) La stimulation de La forêt aléatoire

((b)) La stimulation de la Régression logistique

((c)) La stimulation de l'arbre décisionnel

La méthode d'ensachage (Bagging)

FIGURE 5.6 – Courbe ROC de la méthode d'ensachage



((a)) L'ensachage de La forêt aléatoire

((b)) L'ensachage de la Régression logistique

((c)) L'ensachage de l'arbre décisionnel

5.3 Discussion des résultats

Les modèles mis en œuvre ont d'â être décrits et analysés séparément, mais l'objectif principal de cette recherche était d'identifier quel modèle peut être considéré comme le meilleur pour résoudre ce problème. Afin de faire cette déclaration, nous devons vraiment comparer tous les modèles formés en ce qui concerne nos métriques prédéfinies comme la valeur AUC, et la matrice de confusion.

Pour faire des déclarations statistiques sur les performances des algorithmes et les comparer, il était nécessaire de collecter leurs résultats.

Dans l'analyse du taux du churn des clients, il peut être plus couteux de déduire à tort que le client ne cherche pas à donner une réduction générale des prix des services aux clients qui ne prévoient pas de quitter l'entreprise.

Comme la priorité dans notre étude de cas est donnée à l'identification des clients de désabonnement plutôt que de ceux qui ne le font pas, la sensibilité est plus pertinente que la spécificité dans nos résultats.

À partir des mesures de classification calculées dans les tableaux, et aussi la représentation de la courbe ROC, nous avons conclu les résultats suivants :

- Tous les modèles sont fiables et efficaces dans la prévision du churn.
- Avec un AUC (Area undercurve) égal à 0,90 ce qui est relativement proche de 1, un rappel de 88%, et L'exactitude de 85%, une précision de 83% et un score F de 85%, le modèle Forêt aléatoire est plus performante que les autres modèles.
- Le difficile vote et le doux ont le même temps d'exécution et ils exécutent avec les mêmes jeux et donnés, mais il y a une petite différence au niveau des mesures, où le difficile vote est meilleur par rapport d'exactitude, précision, et le rappel, mais le doux vote a la meilleure valeur par rapport l'auc.
- le temps d'exécution des algorithmes de la méthode de stimulation (boostin) est plus long comparativement à celui de la méthode d'ensachage (bagging), et ça c'est techniquement normal car la méthode d'ensachage fonctionne en parallèle et la méthode de stimulation en séquences.
- Pour la méthode de stimulation et la méthode d'ensachage : La forêt aléatoire est le meilleur algorithme par rapport d'exactitude avec valeur de 85%, précision , et le f_mesure (85%), mais la régression logistique a la meilleure valeur de l'auc avec 0.89. Mais si on va prendre en considération le temps d'exécution, la forêt aléatoire est le pire modèle.
- Alors que la régression logistique est un peu inférieure à la forêt aléatoire en ce qui concerne la performance mais elle est très raisonnable en ce qui concerne le temps d'exécution.

5.4 L'Interface d'application

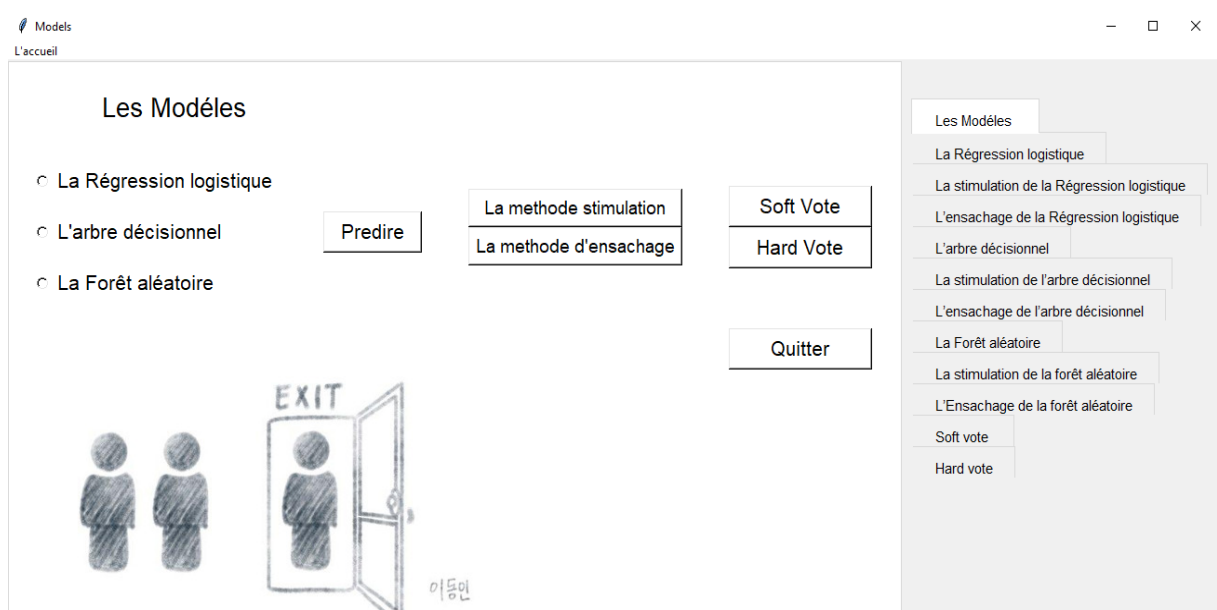
Dans cette partie nous allons détailler les différentes pages de notre interface.

5.4.1 La Page d'accueil

Cet espace permet à l'utilisateur de choisir l'algorithme pour la prédiction. En utilisant les buttons suivants :

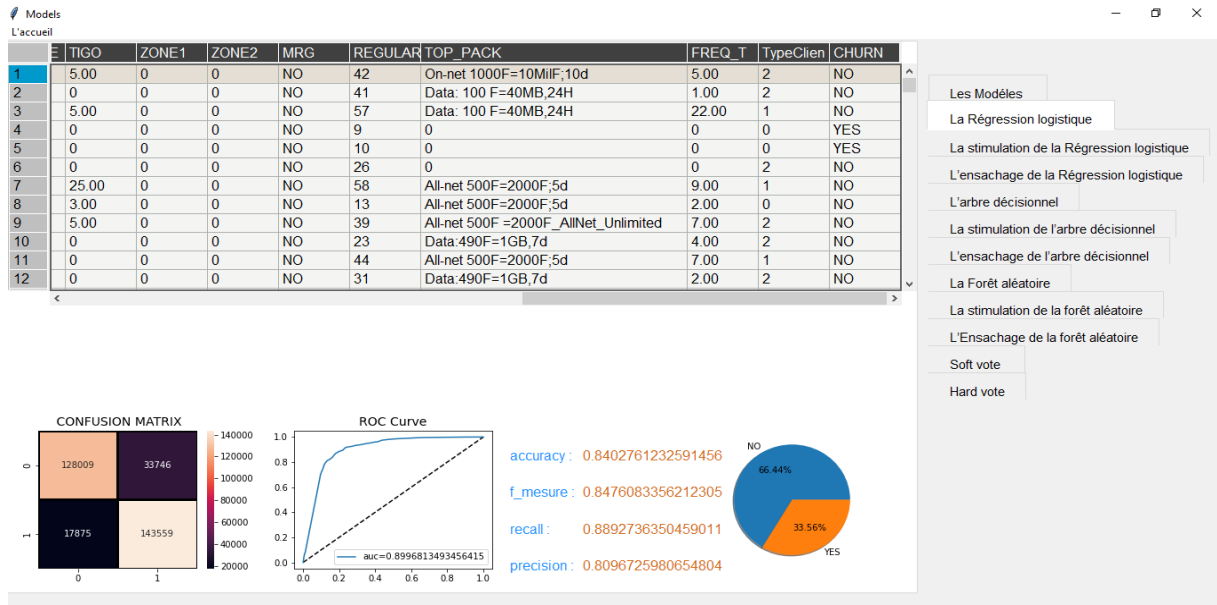
- Prédire : permet de choisir le fichier des clients pour détecter les clients qui vont quitter parmi eux.
- La méthode de stimulation : permet d'appliquer la technique de stimulation sur un algorithme choisi.
- La méthode d'ensachage : permet d'appliquer la technique d'ensachage sur un algorithme choisi.
- Soft Vote : permis d'appliquer le modèle doux vote (soft voting) sur un fichier des clients pour détecter les clients qui vont quitter parmi eux.
- Hard Vote : permis de choisir le fichier des clients pour détecter les clients qui vont quitter parmi eux, en utilisant l'algorithme de difficile vote (hard voting).
- Quitter : pour quitter la fenêtre.

FIGURE 5.7 – La Page d'accueil



5.4.2 La Page des résultats de prédiction

FIGURE 5.8 – La Page des résultats de prédiction



Sur cette page les résultats sont affichés dans un tableau où la colonne churn indiquant si le client va quitter (yes) ou non (no).

Ainsi que les métriques de validation de modèle choisi :

La matrice de confusion, La courbe des caractéristiques d'exploitation du récepteur, Le taux de réussite, Le rappel, La Précision, Le F_mesure.

Et un diagramme circulaire qui indique le pourcentage des clients qui sont restés et les clients qui sont partis.

Il y a une fenêtre pour les résultats de chaque de modèle.

5.5 Outils et technologies utilisées

Les techniques adéquates qu'on a utilisées sont :

5.5.1 Environnement de travail

Python

Python est un langage de programmation interprété de haut niveau, polyvalent et il est le langage le plus utilisé en data science et en machine learning.

Visual Studio Code

Visual Studio Code est un éditeur de code source léger mais puissant qui s'exécute sur votre bureau et est disponible pour Windows, macOS et Linux. Il est livré avec un support intégré pour JavaScript, TypeScript et Node.js et dispose d'un riche écosystème d'extensions pour d'autres langages (tels que C ++, C , Java, Python, PHP, Go) et des runtimes (tels que .NET et Unity).[57]

Dans ce qui suit nous allons présenter les outils et les packages que nous avons utilisées dans chaque partie de notre travail.

5.5.2 Analyse exploratoire des données

Pandas

Pandas est une librairie écrite en Python pour la manipulation et l'analyse des données. Il propose en particulier des structures de données et des méthodes pour manipuler des tableaux numériques et des séries temporelles.

Il a été utilisé dans le nettoyage principalement et pour lire les données en un dataframe qui peut être manipulé, visualisé et transformé facilement.

Numpy

Numpy est la bibliothèque de base pour le calcul scientifique en Python. Il fournit un objet de tableau multidimensionnel hautes performances et des outils pour travailler avec ces tableaux.

Numpy et Pandas seront utilisés dans toutes les parties car ils sont indispensables pour effectuer la majorité des traitements.

Seaborn, Matplotlib, Plotly

Ce sont des bibliothèques du langage de programmation Python destinées à tracer et visualiser des données, des graphiques statistiques et informatiques comme créer des : histogrammes, lineplots, nuage de points, ...etc.

5.5.3 Prédiction

Sklearn

Sklearn ou scikit-learn est un module Python intégrant des algorithmes d'apprentissage automatique classiques dans le monde des packages scientifiques Python (numpy, scipy, matplotlib). Il vise à apporter des solutions simples et efficaces aux problèmes d'apprentissage, accessibles à tous et réutilisables dans divers contextes de machine learning.

Et nous avons utilisé ce module dans la partie pré-traitement de données comme le scaling des données avec MinMax Scaler.

5.5.4 L'Interface graphique

tKinter

tKinter est la bibliothèque graphique libre d'origine pour le langage Python, permettant la création d'interfaces graphiques. Elle vient d'une adaptation de la bibliothèque graphique Tk écrite pour Tcl(Tool Command Language).[58]

5.6 Conclusion

Dans ce chapitre nous avons commencé d'abord par les résultats obtenus des modèles choisis puis les comparer avant et après l'utilisation des techniques d'ensemble, et on observe une augmentation de temps d'exécution avec une amélioration en niveau de la performance pour certains modèles, et une stabilité pour les autres.

Et à la fin nous avons cité le matériel ainsi que les outils et les techniques utilisées dans chaque partie du travail.

Chapitre 6

Conclusion Général

6.1 Conclusion

L'importance de ce type de projet pour le marché des télécommunications ne cesse de croître. La collecte de données devient une tâche quotidienne pour toutes les entreprises, et la valeur de ces données peut provenir de sources multiples.

Ce projet de fin d'études vise à déterminer le modèle approprié qui prédit le taux du churn des clients. Ces modèles devaient enregistrer des valeurs élevées dans les métriques définies : Précision, AUC, rappel, mesure et taux de réussite.

Nous avons tout d'abord réalisé une étude préliminaire sur les jeux de données, toutes les opérations de prétraitement des données telles que le nettoyage des données et la manipulation des valeurs manquantes ont été appliquées, après nous avons éliminé les variables qui n'avaient pas de relations avec la variable cible.

Dans le but d'obtenir les meilleurs résultats possibles, la phase de prétraitement avait pris 50% de notre temps pour élaborer ce projet, cette phase est primordiale pour tout projet d'exploration de données.

Après avoir "nettoyé" le jeu de données, nous avons effectué un regroupement des clients, avec l'algorithme K_Means au niveau de la régularité et qui nous donne 3 types de client : très actif, actif, et inactif.

Puis nous avons procédé à la phase de prédiction. Pour ce faire, nous avons utilisé plusieurs méthodes dues data mining comme la régression logistique, Arbre de décision, forêt aléatoire, et nous avons utilisé les méthodes d'ensemble de stimulation, d'ensachage, et de vote.

Dans le but de construire ces modèles prédictifs, nous avons divisé l'ensemble de données du pilote en deux ensembles de données : Train (60%) et Test (40%).

Avec nos propositions dans ce projet, nous avons pu prouver que les modèles de classification appliqués à l'ensemble de données disponibles, correspondent bien et ils ont produit de bons modèles pour prédire le taux de churn des clients.

6.2 Perspectives

Bien que nous considérons l'étude menée comme un succès, nous reconnaissons qu'il y a encore place à des améliorations.

L'ensemble de données disponible pour cette recherche a une grande taille en ce qui concerne le nombre d'entrées. Cependant, le nombre de variables disponibles à prévoir est assez limité et nous n'avons aucune information démographique considérée comme importante pour ce genre de prédiction.

En raison de la protection des renseignements personnels des employés, l'ensemble de données disponible ne contient pas les données personnelles (le nom, prénom, l'âge, et les dates des appels les durées des appels). Il serait très intéressant d'ajouter ces variables à nos données et de voir si elles améliorent les modèles, et peut-être utilisé de deux façons pour améliorer notre recherche :

Utiliser toutes les données pour prévoir et former les modèles.

Utiliser une segmentation clientèle RFM (récence, fréquence, montant), ou FRAT (la fréquence, la récence, le montant et la catégorie des produits achetés).

Bibliographie

- [1] MARCO RICHELDI AND ALESSANDRO PERRUCCI. Churn Analysis Case Study. Telecom Italia Lab. December 17, 2002.
- [2] JAE-HYEON AHN, SANG-PIL HAN, YUNG-SEOP LEE. Customer churn analysis : Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. 2006 .
- [3] FREDERICK F. REICHHELD. The Loyalty Effect. 1996.
- [4] A. DJEDDOU. Implementation d’une solution de gestion de la relation client (crm),est –elle une tache anodine ?.
- [5] A. Meier. Customer relationship management, 2008.
- [6] “The Evolution of Customer Relationship Management System.” Proceedings of the 19th International Conference on Computers, 2015, p.30.
- [7] KENNEDY (A) : « Electronic Customer Relationship Management (E_CRM) : Opportunities and Challenges in a Digital World.» Irish Marketing Review, 2006, 18(1). P.63.
- [8] JELINEK (D) “The Evolution of Customer Relationship Management System.” Proceedings of the 19th International Conference on Computers, 2015, p.30.
- [9] KIM-SOON (N), ZULKIFLI (M. F) : « The impact of E_CRM Business performance of small company.» Journal of Engineering and Technology 3, p.145.
- [10] CHANDRA (S), STRICKLAND (T. J) : « Technological Differences Between CRM E_CRM.» Issues in Information Systems, 5(2),2004, 408–413.
- [11] PAN (S. L), LEE (J. N) : « Using E_CRM for a unified view of the customer,» Communications of the ACM, 46(4), 2004, p.95–96.
- [12] SIMON F.X, SOUSA M.D. Management et gestion d’un point de vente. Dunod,Paris, 2008

- [13] Maha RAJAB, Gilbert FARGES, Alvin PANJETA . ICARE : POUR AMELIORER LA SATISFACTION DES CLIENTS INTERNES DE L'ADMINISTRATION PUBLIQUE. Université de Technologie de Compiègne, page 141, 2015.
- [14] C. BARBARAY. Satisfaction, fidelite et experience client. Dunod, Paris, 2016.
- [15] <https://www.paperblog.fr/7632162/les-19-impacts-de-la-satisfaction-clients-sur-la-rentabilite-des-entreprises/>.
- [16] R. MATTISON. THE TELCO CHURN MANAGEMENT HANDBOOK. Brigitte Kilger Mattison, 2001.
- [17] Manpreet Kaur, Dr. Prerna Mahajan. (2015). Churn Prediction in Telecom Industry Using R. International Journal of Engineering Research and Applications (IJERA).
- [18] <https://blog.2checkout.com/keep-voluntary-churn-at-minimum/>.
- [19] Essam Shaaban, Yehia Helmy, Ayman Khedr, Mona Nasr. (2012). A Proposed Churn Prediction Model. International Journal of Engineering Research and Applications (IJERA). 2.693-697.
- [20] <https://www.semanticscholar.org/paper/Modeling-%26-Simulation-of-a-Predictive-Customer-for-AwoyeluI./6f41c3d1dbd05567fa8b3e455bad5343647e94ed>.
- [21] Marius Baraitaru. Keep Voluntary Churn at a Minimum and Protect Your Bottom Line. 2020.
- [22] Jae-Hyeon Han, Sang Pil Lee, Yung-Seop. (2006). Customer churn analysis : Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecommunications Policy. 30. 552-568.
- [23] D., Jainam Shah, Fenil Rahevar, Mrugendrasinh. (2018). Customer Churn Prediction Analysis. International Journal of Computer Applications. 182. 15-17.
- [24] Aurélien Géron. Hands-On Machine Learning with Scikit-Learn and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, 2017.
- [25] <https://businesspartnermagazine.com/how-machine-learning-working-step-by-step/>.
- [26] Jason Bell. Machine Learning : Hands-On for Developers and Technical Professionals. Indianapolis, Indiana. 2015.
- [27] Ayodele, T. O. (2010). Types of machine learning algorithms. New advances in machine learning, 3 :19-48.

- [28] Dudarev, S., Botton, G., Savrasov, S., Humphreys, C., and Sutton, A. (1998). Electron-energy-loss spectra and the structural stability of nickel oxide : An lsd+u study. *Physical Review B*, 57(3) :1505.
- [29] Chang, R.-I., Lai, L.-B., Su, W., Wang, J.-C., and Kouh, J.-S. (2006). Intrusion Detection by Backpropagation Neural Networks with Sample-Query and Attribute-Query, pages 6–10. Research India Publications.
- [30] Insights, R. Machine learning for fraud detection.
- [31] Patidar, R. and Sharma, L. (2011). Credit Card Fraud Detection Using Neural Network, volume 1, pages 32–38. *International Journal of Soft Computing and Engineering (IJSCE)*.
- [32] Ngai et al. (2011). The application of data mining techniques in financial fraud detection : A classification framework and an f review of literature.
- [33] Kirkos et al. (2007). Data mining techniques for the detection of fraudulent financial statements
- [34] Alkhateeb, Z. K. and Maolood, A. T. (2019). Machine learning-based detection of credit card fraud : A comparative study.
- [35] Miroslav Kubat. *An Introduction to Machine Learning Second Edition*, Springer Nature, 2015
- [36] Bolton, R. J. and Hand, D. J. (2001). Unsupervised profiling methods for fraud detection
- [37] Lata, L. N., Koushika, I. A., and Hasan, S. S. (2015). A comprehensive survey of fraud detection techniques.
- [38] Aswathy M S, L. S. (2018). Survey on credit card fraud detection.
- [39] Introduction to K_means Clustering (oracle.com)
- [40] Ait Mahammed, F. (2018). Approches d'apprentissage automatique pour la détection du spam web : exploration de diverses caractéristiques.
- [41] https://openclassrooms.com/fr/courses/4525281-realisez-une-analyse-exploratoire-de-donnees/5177935-decouvrez-l-algorithme-k_means.
- [42] https://eric.univ-lyon2.fr/ricco/cours/slides/classif_centres_mobiles.
- [43] Zaki, M. J. and Meira, M. J. (2013). *Data Mining and Analysis : Fundamental Concepts and Algorithms*

- [44] R.Kumari, Sheetanshu, and M.K.Singh (2016). Anomaly detection in network traffic using k-mean clustering.
- [45] https://fr.abcdef.wiki/wiki/Oversampling_and_undersampling_in_data_analysis.
- [46] <https://www.pinterest.com/pin/514958538641697615/>.
- [47] Jason Brownlee. (2020).Train_Test Split for Evaluating Machine Learning Algorithms.
- [48] Freund, Y.et Schapire, R.E. Experiments with a new boosting algorithm. The 13th International Conference on Machine Learning.pp.148-156.1996.
- [49] <https://machine-learning.paperspace.com/wiki/gradient-boosting>.
- [50] <https://fr.acervolima.com/ml-classificateur-de-vote-utilisant-sklearn/>
- [51] <https://towardsdatascience.com/types-of-ensemble-methods-in-machine-learning-4ddaf73879db>.
- [52] <https://www.datacamp.com/community/tutorials/adaboost-classifier-python>.
- [53] https://en.wikipedia.org/wiki/Bootstrap_aggregating.
- [54] Rahm, E. and Do, H. (2000). Data cleaning : Problems and current proaches. IEEE Data Engineering Bulletin, 23 :3–13.
- [55] <https://medium.com/@natratanonkanraweekultana/confusion-matrix-d6146b275faa>.
- [56] https://www.researchgate.net/figure/Receiver-operating-characteristic-ROC-curves-with-respective-area-under-the-curve-AUC_fig3_331311273.
- [57] <https://code.visualstudio.com/learn>
- [58] <https://wiki.python.org/moin/TkInter>