Prediction des clients susceptibles de retirer leurs fonds

Mohammed BOURI
Mohammed EL ATTAR SOFI

2019/2020

Résumé

L'objectif principal de cette étude serait d'effectuer des analyses et de trouver des potentiels consignes pour les preneurs de décision, concernant les différentes stratégies possibles pour la prévention du Churn. Le fait qu'un client décide de quitter une banque représente un affaiblissement considérable de revenus. Nous avons mené une démarche consistant sur l'exploitation des différentes techniques de Data Mining et de Machine Learning afin d'avoir une idée approximative sur le comportement du client et s'il est susceptible de quitter l'établissement.

Pour assurer une organisation de travail optimale, nous avons choisi d'adopter Jupyter Notebook comme environnement de développement, cela nous a permis de combiner le code du logiciel, les résultats de calcul et le texte explicatif en une seule entité. Nous avons commencé par une exploration approfondie des deux bases de données à notre disposition pour une meilleure compréhension de la situation des clients échantillonnés. Afin d'assurer une prévision significative de la variable Churn, nous avons consulté plusieurs études théoriques discutant ce phénomène, nous avons alors repéré les variables nécessaires afin de mener notre prévision. Une fois les variables transformés et optimisés, nous avons effectué le nettoyage de données qui est une étape cruciale pour préserver une performance algorithmique maximale. L'entrainement du modèle nous a alors donné des résultats satisfaisants en termes de précision et de significativité.

Ce modèle construit est capable de différencier les personnes susceptibles de quitter la banque en fonction de ces comportements. Ceci aura possiblement un rôle important dans l'aspect marketing de la banque, surtout au niveau du ciblage. Une fois les modèles entrainés, la banque pourra détecter les attributs qui caractérisent les personnes qui pourront churner. Ceci fait, des stratégies préventives sont alors mises en place afin de convaincre les churners de revenir et pour éviter le départ d'autres clients.

Table des matières

Liste des figures

Liste des tableaux

Première partie

Introduction générale

Introduction

Avant de discuter l'essence de notre étude, contextualisons d'abord notre sujet. Avant la crise financière de 2008, les banques étaient déterminées à investir dans l'élargissement inconditionnel du nombre des clients. L'effondrement du marché a par contre transformé la stratégie du secteur bancaire envers leur clients. Il s'est avéré que le gain de nouveaux clients coûtait sept fois plus que la préservation des clients déjà existants, la perte de ces clients causerait alors des dommages financiers significatifs. Cette possibilité est de plus en plus réalisable dernièrement, puisqu'il est dorénavant plus facile de déplacer des actifs et de l'argent entre les différentes banques. La politique concurrentielle menée par les banques offre au client plusieurs choix variés, il peut par conséquent migrer d'une banque à une autre du jour au lendemain.

Les banques, compagnies d'assurance et compagnies de télécommunication utilisent très fréquemment les méthodes de prévention de « Churn », qui signifie simplement la perte des clients, en menant une analyse sur les comportements des clients qui ont quitté l'organisme et exploiter ces ressources afin de réaliser des prévisions. Celles-ci nous permettront de déterminer les clients les plus potentiellement susceptibles de quitter l'entité en question. Des stratégies de prévention sont alors mises en place en se basant sur les résultats des prévisions.

Ceci est rendu possible bien plus que jamais grâce aux différents algorithmes de Machine Learning qui maximisent le plus réellement possible le succès et la significativité des prévisions et entraînent alors de meilleures décisions.

Nous allons alors exploiter nos différentes connaissances acquis lors de notre parcours académique, et mener plusieurs recherches approfondies afin de pouvoir construire un modèle qui pourra nous révéler les clients qui pourront potentiellement quitter une entité particulière.

Deuxième partie

Cadre théorique

Chapitre 1

Théorie de l'étude

1.1 Data Science, Big Data et Machine Learning

La science des données (Data Science) représente le traitement absolu de l'information, en ce qui concerne son origine, ce qu'elle constitue et comment nous pourrions l'exploiter comme une richesse importante dans les différentes prises de décisions, qui figurent dans les politiques de marketing et de gestion. En effet, l'importance de cette discipline est clairement perçue, pour une entreprise quelconque, dans sa contribution dans l'accroissement de la performance et l'efficacité des stratégies mises en place, le contrôle effectif des coûts et la détection rapide d'opportunités novatrices de marché pouvant améliorer les chances concurrentielles de l'entreprise en question. Les tendances identifiées ne peuvent être obtenues que grâce à l'extraction de grandes quantités de données. L'introduction du terme Big Data s'avère alors nécessaire.

Cette discipline représente les quantités énormes de données structurées, semi-structurées et non structurées pouvant être exploités. Quand nous parlons de Big Data, nous insinuons un volume extrême de données, une grande variété de types de données et une vitesse supérieure à laquelle les données doivent être traitées.

Quant au Machine Learning, c'est une discipline qui permet l'utilisation et le développement de systèmes informatiques capables d'apprendre et de s'adapter sans suivre des instructions explicites, en utilisant des algorithmes et des modèles statistiques pour analyser et tirer des inférences à partir des modèles de données. Les modèles construits servent principalement à la réalisation de prévisions et d'estimations essentielles pour la survie de l'entreprise.

1.2 Le « Churn » des clients

Le « Churn » est quand un client existant, un utilisateur, un joueur, un abonné ou tout autre type de client, cesse volontairement de jouir des services d'une entreprise ou met fin à sa relation avec elle.

Cela peut signifier l'annulation d'un abonnement, la fin d'une adhésion ou la fermeture d'un compte d'une manière ou d'une autre. Il peut également s'agir de ne pas renouveler un contrat ou un accord de service et même d'un client qui décide d'acheter ses courses dans un autre magasin.

La perte de clientèle est une réalité angoissante qui touche toutes les entreprises à un moment ou à un autre. Même les entreprises les plus grandes ou les plus prospères ne sont pas épargnées par la défection des clients. Il est important, pour une croissance durable et viable des entreprises, de comprendre ce qui peut pousser des clients et des utilisateurs auparavant fidèles à abandonner le navire et à trouver une nouvelle entité.

1.3 Data Mining

1.3.1 Définition du Data Mining

En se basant sur des bases de données de taille importante, le Data Mining nous permet l'extraction d'informations prédictives dans ces BDs. Son objectif majeur est d'effectuer des recherches avancées afin de repérer les informations les plus nécessaires. Nous posséderons par conséquent des modèles antérieurement anonymes et des informations additionnelles. Ce gain pourrait être exploité dans les prises de décision adéquates aux recherches effectuées.

Le processus suit généralement les étapes suivantes :

- Les données doivent, avant toute chose, être explorées et optimisés. En effet, compte tenu du nombre massif des données à disposition, les erreurs de saisie et les imperfections sont inévitables. Les données qui apparaissent plus qu'une fois ou qui ne concernant pas notre étude doivent alors être éliminés. Des transformations peuvent aussi avoir lieu si cela se montre nécessaire.
- Une fois notre base de données renouvelée, nous pouvons entamer l'étape du repérage des modèles que nous allons créer. Il est impératif que ces modèles soient convenables

aux données choisies et qu'ils inscrivent les meilleures prévisions possibles.

- L'étape finale est la mise en marche des modèles choisis et les appliquer. L'aboutissement des prévisions doit nous rapprocher du but original.

1.3.2 Le Data Mining dans le secteur bancaire

Les données enregistrées dans le secteur bancaire sont d'une quantité colossale, elles sont constituées des données relatives aux transactions : La date, le nombre et le montant des transactions liés aux identifiants correspondants. Ces données donnent aussi des informations sur les clients : Sa banque, son agence, sa ville, son sexe, son âge et encore plus... Il est alors impossible de traiter manuellement ces données, d'où l'importance du Data Mining qui permet de mieux identifier sa clientèle et interagir avec celle-ci quand c'est nécessaire. Il permet par exemple le processus de détection de fraude dans le secteur bancaire, l'identification des comportements anormaux et le contrôle du risque du vol des données personnelles.

Ce secteur est le leader en termes de volume données, bien plus que les réseaux sociaux. Nous ferons alors face à plusieurs complications dans le traitement des données. Le premier reste évident, les données enregistrées peuvent atteindre des milliards et leur traitement efficace pourrait se montrer compliqué vu la grande échelle dont elles appartiennent.

Le second problème est qu'on retrouve que ces données sont brutes et ne peuvent pas être exploitées instantanément dans l'exploration des données. Hormis les problèmes de redondance et de valeurs manquantes, nous devrons impérativement créer de nouvelles variables en nous basant sur les variables déjà existantes. Ces nouvelles variables seront récapitulatives des informations dont nous aurons besoin pour travailler.

Les phénomènes traités dans ce cadre sont qualifiés de très rares, notre sujet d'étude (« Churn ») en est la plus grande preuve. Ceci est le cas également pour les projets portant sur la détection des fraudes et des attitudes malhonnêtes. La rareté est alors une autre complication qui doit être étudiée.

Ce processus a donc logiquement une immense contribution pour les stratégies de marketing bancaire. En effet, les clients sont aisément identifiés et il est alors possible d'avoir une idée claire sur leurs profils et, en se basant sur cela, générer des stratégies de marketing adaptées aux caractéristiques des clients. L'organisme pourra alors prendre efficacement des décisions et de mieux contrôler le rendement et les résultats.

1.4 Modèles prédictifs

Dans cette étude, nous aurons besoin de construire un modèle afin de prévoir notre variable cible. Nous nous baserons sur les données historiques qui sont à notre disposition afin de générer notre modèle et le valider. Généralement, la présence de variables indépendantes est le facteur qui peut potentiellement agir sur le comportement futur.

Il existe deux cas de modélisation prédictive :

- Si la variable cible est qualitative, la modélisation représente une classification
- Si la variable cible est quantitative, la modélisation représente une régression

1.4.1 Régression

Afin d'estimer une valeur numérique en sortie, la régression s'avère efficace. Elle nous permet d'exhiber le lien entre les différentes variables dépendantes et indépendantes de notre base de données. Nous pourrions l'exploiter par exemple afin d'estimer la température (Variable quantitative) d'une zone précise en fonction de sa latitude et sa longitude. Dans ce même exemple, nous pourrions intégrer une variable qui décrit le temps, nous aurions affaire à une régression avec des données temporelles. La variable cible représenterait alors la température, le reste des variables représenterait les prédicteurs.

L'algorithme de régression choisi se chargera alors d'estimer la variable cible évaluée en fonction des prédicteurs. Cette étape comporte plusieurs itérations, les valeurs originales que nous avions utilisé pour la mise en place du modèle sont éliminées, et nous pourrons par conséquent appliquer ce modèle résultant sur des données différentes et avoir les résultats souhaités.

Il est important de souligner enfin que nous parlons de modèle de régression linéaire lorsque les paramètres du problème peuvent former une combinaison linéaire instantanément ou grâce à des transformations que nous pouvons entamer comme la technique de changement de variables. Par contre, il est impossible parfois, en fonction des données mises en jeu, d'obtenir un modèle linéaire. Nous recourrons alors à des algorithmes différents appartenant, par exemple, aux estimations non paramétriques. Il nous faut enfin vérifier la qualité de notre régression et si elle est la plus convenable à nos données.

1.4.2 Classification

La classification est une méthode qui nous permet de prévoir un résultat, comme pour la régression, mais concerne principalement les variables cibles de type catégoriel. La génération du modèle s'effectue en fonctions de variables numériques et qualitatives.

Il est nécessaire que nous ayons à notre disposition des données initiales décrivant les valeurs des prédicteurs et de la variable cible. Le processus de classification se charge alors de rechercher les liaisons entre les différents attributs qui vont nous permettre de trouver notre prévision. Une fois ceci établi, nous pourrions appliquer notre modèle de classification sur de nouvelles données totalement anonymes pour l'algorithme mais à condition qu'on ait les mêmes variables que ceux avec quoi on a construit notre modèle en ôtant la variable cible pendant ce processus. Une fois que l'algorithme est complété, la prévision est obtenue en fonction des données récentes. L'étape finale est le jugement de la performance du modèle en question et sa validation à travers des techniques adaptés à notre situation.

Puisque l'étude de la prévention du Churn se base une variable cible de type qualitatif, nous procéderons dans ce projet aux différentes méthodes de classification.

1.5 Clustering

Le principe est assez simple : Nous regroupons les observations qui ont des degrés de similitude en un seul cluster homogène. Par conséquent, les observations appartenant à d'autres groupes auront des caractéristiques différentes.

Il existe plusieurs algorithmes capables d'effectuer le partitionnement des données, on choisit alors le meilleur par rapport aux attributs de l'ensemble des données à notre disposition.

1.6 Revue des méthodes utilisés

1.6.1 La méthode d'arbre de décision

La technique d'arbre de décision, appartenant aux techniques d'apprentissage supervisé, s'effectue en distinguant étape par étape l'ensemble formé pour l'entraînement dans le but de générer des groupes aussi purs que possible pour une classe cible déterminée.

Théoriquement, le processus relie chaque nœud appartenant aux arbres à un nombre optimal d'observations T. Celles-ci sont distingués grâce à un test récursif sur un attribut prédéterminé. - La distinction pour une caractéristique quantitative A est trouvée par le test $A \leq x$. La totalité des observations T est alors divisé en deux groupes qui se dirigent à la branche droite et gauche de l'arbre.

$$T_g = \{t \in T : t(A) \leq x\} \qquad T_d = \{t \in T : t(A) > x\}$$

On procède approximativement de la même manière pour une variable qualitative qu'on nomme B. Supposons que $B = \{b_1, ..., b_k\}$, alors le test $B = b_i$ nous aide à déterminer la branche i.

Le modèle résultant est employé afin de réaliser une prévision sur la période où le client est toujours souscris à une banque et ne l'a pas encore quitté, en fonction des données démographiques relatives à chaque client.

1.6.2 La méthode de régression logistique

La classification peut être effectuée grâce à la méthode de régression logistique. Elle appartient aux techniques d'analyse en présence de plusieurs variables. Son importance réside dans l'estimation de la liaison entre l'émergence d'un phénomène qui sera de type qualitatif et les éléments qui pourraient potentiellement agir sur lui. En d'autres termes, ces éléments représentent les variables explicatives du problème.

Ses champs d'application sont immenses, nous pourrions citer son rôle dans le monde des assurances où elle alloue la détermination de la partie des clients susceptibles à souscrire à une assurance pour couvrir un risque déterminé. Les chercheurs en médecine l'emploient aussi pour détecter les facteurs qui distinguent un ensemble de gens malades des autres qui ne souffrent pas d'une maladie.

Par rapport à notre domaine d'étude qui est le secteur bancaire, cette méthode de régression logistique est valable pour trouver les personnes susceptibles de payer leur crédit à échéance ou non et aussi pour repérer les clients qui peuvent quitter définitivement la banque.

Afin de mieux comprendre le fonctionnement pratique de cette méthode, essayons d'abord d'assimiler son aspect théorique :

Supposons qu'on dispose de la probabilité estimée (notée Y) qu'un client quitte une banque précise. Nous posons $u = B_0 + B_1X_1 + B_2X_2 + ... + B_kX_k$ qui décrit la régression linéaire effectuée avec $B_0, B_1, B_2, ..., B_k$ qui représentent des valeurs constantes. L'ensemble des variables $X_1, X_2, ..., X_k$ constituent les variables indépendantes du problème. La définition de Y s'écrit de la façon suivante :

$$Y = \frac{e^u}{1 + e^u}$$

Dans le cas de notre étude, la régression logistique pourrait être employé pour analyser les conséquences des éléments socio-économiques sur le « Churn » des clients en se basant sur ceux qui sont les plus susceptibles de quitter la banque. Le but principal est d'analyser les effets des attributs socio-économiques et démographiques sur la possibilité le départ d'un client ou d'un autre.

Dans ce cas, il s'est avéré que presque la totalité des variables indépendantes sont extrêmement nécessaires pour la réalisation de la prévision.

1.6.3 Méthode du Random Forest

Une grande partie du Machine Learning est la classification - nous voulons savoir à quelle classe (ou groupe) appartient une observation. La capacité de classer précisément les observations est extrêmement précieuse pour diverses applications commerciales, comme la prévision de l'achat d'un produit par un utilisateur particulier ou la prévision de la défaillance ou non d'un prêt donné.

Le Random Forest est un algorithme d'apprentissage supervisé qui fait partie des méthodes d'ensembles. Le principe des méthodes d'ensemble est basé sur le regroupement de plusieurs algorithmes d'apprentissage instables pour créer un algorithme plus performant.

Le Random Forest est le regroupement de plusieurs arbres de décision, où chaque arbre est construit sur un échantillon tiré avec remise de la base de données d'apprentissage, avec un choix aléatoire des variables de séparation. La prédiction du modèle est la moyenne des prédictions de l'ensemble des arbres de décisions dans le cas de la régression, et le vote majoritaire dans le cas de la classification.

La performance du Random Forest vient du regroupement de plusieurs arbres de décision, qui permet de réduire la variabilité dans la variable sortie, et le choix aléatoire des variables de séparation qui permet de décorréler les arbres de décision construites, et par conséquent réduire l'erreur final.

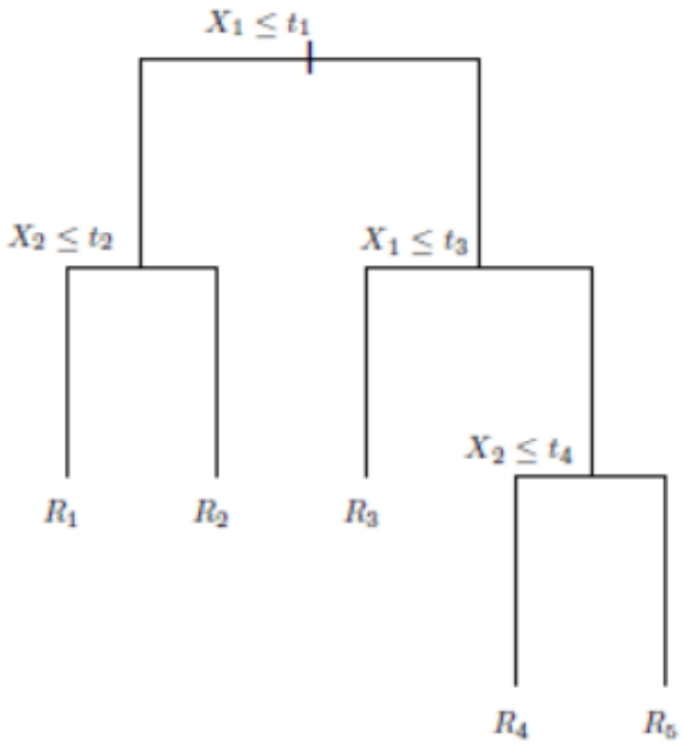

FIGURE 1.1 – Exemple d'arbre de decision pour deux variables

Troisième partie

Cadre pratique

Chapitre 2

Outils choisis

2.1 Jupyter

Jupyter est un outil web interactif, gratuit et à code source ouvert, connu sous le nom de carnet de calcul, que les chercheurs peuvent utiliser pour combiner le code logiciel, les résultats de calcul, le texte explicatif et les ressources multimédia dans un seul document. Les carnets de notes informatiques existent depuis des décennies, mais Jupyter en particulier a explosé en popularité au cours des dernières années. Cette rapide adoption a été favorisée par une communauté enthousiaste d'utilisateurs-développeurs et une architecture redéfinie qui permet au bloc-notes de parler des dizaines de langages de programmation. Cette croissance est due à l'amélioration des logiciels web qui pilotent des applications telles que Gmail et Google Docs, à la maturation du python scientifique et de la science des données, et, surtout, à l'aisance avec laquelle les ordinateurs portables facilitent l'accès à des données à distance qu'il serait autrement peu pratique de télécharger.

Pour les spécialistes des données, ce format peut stimuler l'exploration. Les ordinateurs portables, sont une forme d'informatique interactive, un environnement dans lequel les utilisateurs exécutent des codes, voient ce qui se passe, modifient et répètent dans une sorte de conversation itérative entre le chercheur et les données. Le Notebook de Jupyter a deux composantes. Les utilisateurs saisissent un code de programmation ou du texte dans des cellules rectangulaires d'une page web frontale. Le navigateur transmet ensuite ce code à un "noyau", qui exécute le code et renvoie les résultats.

Ces outils favorisent la reproductibilité des calculs en simplifiant la réutilisation des codes. Mais les utilisateurs doivent encore savoir comment utiliser correctement ces Notebooks.

2.2 Python

Les chiffres ne mentent pas. Selon des études récentes, Python est le langage de programmation préféré des scientifiques de données. Ils ont besoin d'un langage facile à utiliser, avec une disponibilité de bibliothèque décente et une grande participation de la communauté. Les projets qui ont des communautés inactives sont généralement moins susceptibles de maintenir ou de mettre à jour leurs plateformes, ce qui n'est pas le cas de Python.

Qu'est-ce qui rend Python si idéal pour la science des données ? Nous avons examiné pourquoi Python est si répandu dans l'industrie florissante de la science des données - et comment vous pouvez l'utiliser pour vos grands projets de données et d'apprentissage de la machine.

Python est connu depuis longtemps comme un langage de programmation simple à reprendre, du point de vue de la syntaxe en tout cas. Python a également une communauté active avec un vaste choix de bibliothèques et de ressources. Le résultat ? Nous disposons d'une plate-forme de programmation qu'il est logique d'utiliser avec les technologies émergentes comme l'apprentissage machine et la science des données.

Les professionnels qui travaillent avec des applications en sciences des données ne veulent pas s'embourber dans des exigences de programmation compliquées. Ils veulent utiliser des langages de programmation comme Python pour effectuer des tâches sans problèmes.

La science des données consiste à extrapoler des informations utiles à partir d'énormes réserves de statistiques, de registres et de données. Ces données sont généralement non triées et difficiles à corréler avec une précision significative. L'apprentissage machine peut établir des connexions entre des ensembles de données disparates, mais il nécessite une grande puissance de calcul.

Python répond à ce besoin en étant un langage de programmation polyvalent. Il permet de créer des sorties CSV pour faciliter la lecture des données dans un tableur. Il est également possible d'obtenir des sorties de fichiers plus complexes qui peuvent être ingérées par des groupes de Machine Learning pour le calcul.

2.3 Bibliothèque Pandas

Pandas est une bibliothèque Python qui est un outil simple mais puissant de la science des données. Python Pandas est l'un des paquets les plus utilisés en Python. Ce paquet comprend de nombreuses structures de données et des outils pour manipuler et analyser efficacement les données. Python Pandas est utilisé dans des domaines tels que l'économie, la comptabilité, l'analyse, les statistiques, etc. partout, y compris dans les secteurs commerciaux et universitaires.

Nous avons choisi d'utiliser la bibliothèque Pandas en Python pour suivre, évaluer et purifier les données. Python Pandas est bien adapté aux données de différentes formes, telles que :

- Les données tabulaires dont les colonnes sont tapées de manière hétérogène
- Les données sur les séries chronologiques ordonnées et non ordonnées
- Les données matricielles
- Les données non identifiées
- Toute autre forme d'ensemble de données statistiques ou d'observations

2.4 Bibliothèque Scikit-learning

Scikit-learn est une bibliothèque du « Machine Learning » qui existe dans « Python » gratuitement. Elle propose divers algorithmes de classification, de régression et de regroupement, notamment des machines à vecteurs de support, Random Forest, ect...

Elle est conçue pour interagir avec les bibliothèques numériques et scientifiques de Python, tel que : NumPy et SciPy.

Chapitre 3

Modélisation prédictive

3.1 Présentation des données

Avant d'exposer les différents composants de notre base de données, préparons d'abord l'environnement où nous allons travailler. Afin de mieux structurer notre tâche de programmation, nous avons choisi d'utiliser Jupyter Notebook et le langage de programmation Python pour coder.

Ceci aussitôt établi, nous devons désormais importer les différentes bibliothèques que nous allons exploiter. Ces bibliothèques sont :

- **Pandas :** Pour le traitement des bases de données
- **Numpy :** pour l'insertion des grands tableaux et matrices multidimensionnels, ainsi qu'une vaste collection de fonctions mathématiques de haut niveau pour opérer sur ces tableaux.

L'organisme où nous avons effectué notre stage d'application nous a confié 2 bases de données différentes afin de mener notre étude sur le phénomène du Churn.

La première base de données contient 287115 lignes et 21 variables. Elle concerne les informations stockées sur les 200 000 clients extraits comme échantillon pour notre étude, à savoir :

- **Banque :** La banque où le client est souscrit
- **Agence :** L'agence que le client consulte pour effectuer des transactions ou pour bénéficier des services

- **Ville :** La ville où le client habite (A vrai dire, c'est la ville où le client a créé son compte bancaire)
- **flag_etranger_res_maroc :** S'il est résident au Maroc ou non
- **flag_proprietaire_logement :** S'il dispose d'un logement ou non
- **secteur_activite :** Le domaine où il exerce son métier
- **flag_app_mobile :** S'il utilise l'application mobile de la banque ou non
- **sexe :** Le sexe du client
- **date_naissance :** Sa date de naissance
- **salary :** Le salaire que le client a mentionné
- **profession :** Le métier du client
- **nombre_enfant :** Le nombre d'enfants du client
- **marital_status :** Le statut marital du client
- **date_eer** La date où le client a créé son compte
- **id :** L'identifiant du client (A noter que chaque client dispose d'un identifiant spécifique à lui-même)

```
Entrée [53]: client.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 287115 entries, 0 to 287114
Data columns (total 21 columns):
 #   Column                       Non-Null Count   Dtype
---  ------                       --------------   -----
 0   banque                       287115 non-null  int64
 1   agence                       287115 non-null  int64
 2   generic                      287115 non-null  int64
 3   plural                       287115 non-null  int64
 4   ccle                         287115 non-null  int64
 5   pnb                          274032 non-null  object
 6   code_ville                   287115 non-null  int64
 7   ville                        287115 non-null  object
 8   flag_etranger_res_maroc      287115 non-null  object
 9   flag_proprietaire_logement   287115 non-null  object
 10  secteur_activite             287115 non-null  object
 11  flag_app_mobile              287115 non-null  object
 12  sexe                         285466 non-null  object
 13  date_naissance               286024 non-null  float64
 14  salary                       287115 non-null  object
 15  profession                   44835 non-null   object
 16  nombre_enfant                190292 non-null  float64
 17  marital_status               286024 non-null  object
 18  date_eer                     287115 non-null  object
 19  segment                      287115 non-null  object
 20  id                           287115 non-null  int64
dtypes: float64(2), int64(7), object(12)
memory usage: 46.0+ MB
```

FIGURE 3.1 – La base de données : Clients

La deuxième base de données contient 16857064 lignes et 8 variables. Elle représente les données regroupées concernant les transactions des 200 000 clients choisis. Les variables en jeu sont :

— **ANNEE** : L'année désignée
— **MOIS :** Le mois désigné
— **SEMAINE :** La semaine désignée
— **NBRE_TRANSACTION :** Le nombre de transactions pendant la semaine désignée
— **MONTANT :** Montant des nombres de transactions pendant la semaine désignée
— **SENS_TRANSACTION :** Si c'est un crédit ou un débit
— **DWP :** La date désignée en format « int »
— **Id :** L'identifiant du client

```
Entrée [55]: transaction.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16857064 entries, 0 to 16857063
Data columns (total 8 columns):
 #   Column            Dtype
---  ------            -----
 0   ANNEE             int64
 1   MOIS              int64
 2   SEMAINE           int64
 3   NBRE_TRANSACTION  int64
 4   MONTANT           float64
 5   SENS_TRANSACTION  object
 6   DWP               int64
 7   id                int64
dtypes: float64(1), int64(6), object(1)
memory usage: 1.0+ GB
```

FIGURE 3.2 – La base de données : Transactions

3.2 Préparation des données

Afin de pouvoir générer et appliquer un modèle de Machine Learning, l'une des étapes les plus cruciales est la préparation des données afin de pouvoir les adapter au modèle choisi. Ceci dit, nous devons désormais sélectionner les variables qui auront un poids significatif pour l'accomplissement de notre prévision.

Plusieurs études théoriques ont montré que l'étude de prévision du « Churn » dans le secteur bancaire se fait principalement grâce à des prédicteurs spécifiques. Ils permettent de

mieux résumer l'information que le modèle choisi exploitera et, par conséquent, d'obtenir de meilleures prévisions en termes de significativité.
Ces variables prédicteurs désignent :

3.2.1 Age

Nous pouvons d'abord remarquer l'existence antérieure de la variable âge dans notre base de données initiale. Or, pour des raisons d'adaptation aux modèles du Machine Learning, et afin de donner au modèle beaucoup plus d'informations où il peut s'entrainer, nous avons choisi de diviser notre base de données en 10 parties en fonction de la date des transactions représentée ici par la variable « DWP ». Ensuite, nous avons calculé l'âge des clients pour chaque « DWP » choisi dans la liste choisie des dates.

Une base de donnée résultante est par conséquent générée comportant les variables « DWP », « id » et «AGE»

```
Entrée [6]: df_AGE.head(4)
```

Out[6]:

	DWP	id	AGE
0	201707	87631	60.0
1	201707	120851	56.0
2	201707	28206	49.0
3	201707	144072	51.0

FIGURE 3.3 – La base de données : Age

3.2.2 Région / Ville

Vu les distinctions au niveau économiques et démographiques entre les différentes villes du Maroc, en plus des différences importantes dans le nombre des habitants dans chaque ville, il est significativement nécessaire d'intégrer les informations disponibles concernant l'emplacement des clients dans la construction de notre modèle.

Puisque notre variable est de type « String » qui désigne une chaîne de caractères, nous devons la transformer en variables de type « Dummy Variable », c'est-à-dire une variable à données binaires qui est égale à 0 ou 1. Par exemple, si un client habite à Casablanca, la

valeur de la colonne « Casablanca » correspondante à ce client prendra la valeur de 1, et les valeurs des autres colonnes correspondantes prendront 0.

Par souci de multitude de villes et pour ne pas fournir à l'algorithme de Machine Learning des informations dont il n'a pas besoin, nous avons choisi de ne prendre en considération que les 6 villes les plus redondantes et regrouper les villes restantes dans une colonne nommée « AUTRE ». Nous obtenons une base de données composée des colonnes : « AUTRE », « CASABLANCA », « FES », « MARRAKECH », « RABAT », « SALE », « TANGER » et les identifiants de chaque client.

3.2.3 Statut marital

L'information sur le statut marital, c'est-à-dire s'il est marié, divorcé, célibataire ou veuf, est assez importante dans la génération du modèle puisqu'elle nous donne une idée approximative sur la situation et du caractère futur du client.

	DIVORCED	MARRIED	OTHER	SINGLE	WIDOWED	ID
0	0	1	0	0	0	87631
1	0	1	0	0	0	120851
2	0	1	0	0	0	28206
3	0	1	0	0	0	144072
4	0	0	0	1	0	78120
5	0	0	0	1	0	157648

FIGURE 3.4 – La base de données : Statut marital

3.2.4 Crédit / Débit

Le montant des transactions est évidemment une information nécessaire pour être capable de prévoir si le client va quitter une banque ou non. Mais comment pourrions-nous la présenter à notre modèle de façon qu'il prévoit efficacement et significativement notre « Churn » ?

Nous avons alors mis en place une fonction qui nous permet de calculer les crédits et les débits pour un nombre, que nous pouvons définir, de mois, et ceci est effectué pour chaque client. Cette manœuvre offrira au modèle des informations clés pour émettre des prévisions. Les deux bases de données résultantes sont sous cette forme :

Crédit

id	DWP	CREDIT_6_MONTH_AGO	CREDIT_5_MONTH_AGO	CREDIT_4_MONTH_AGO	CREDIT_3_MONTH_AGO	CREDIT_2_MONTH_AGO	CREDIT_1_MONTH_AGO
631	201707	1000.0	0.0	1000.0	0.0	0.0	2351.0
631	201710	0.0	0.0	2351.0	2351.0	2351.0	2351.0
631	201801	2351.0	2351.0	2351.0	2351.0	2351.0	2351.0
631	201804	2351.0	2351.0	2351.0	2351.0	2351.0	2351.0
631	201807	2351.0	2351.0	2351.0	6351.0	2351.0	2351.0

FIGURE 3.5 – La base de données : Crédit

Débit

id	DWP	DEBIT_6_MONTH_AGO	DEBIT_5_MONTH_AGO	DEBIT_4_MONTH_AGO	DEBIT_3_MONTH_AGO	DEBIT_2_MONTH_AGO	DEBIT_1_MONTH_AGO
87631	201707	-102.0	-215.0	-116.0	-15.0	-15.0	-15.0
87631	201710	-15.0	-15.0	-15.0	-15.0	-15.0	-15.0
87631	201801	-15.0	-15.0	-15.0	-15.0	-15.0	-15.0
87631	201804	-15.0	-15.0	-15.0	-101.0	-15.0	-15.0
87631	201807	-101.0	-15.0	-15.0	-20.0	-15.0	-15.0

FIGURE 3.6 – La base de données : Débit

3.2.5 PNB

Nous introduisons aussi la valeur du PNB qui représente la valeur ajoutée générée par les services de la banque.

	id	pnb
0	87631	1022.21
1	120851	224.41
2	28206	23940.36
3	144072	9.45
4	78120	1393.3

FIGURE 3.7 – La base de données : PNB

3.2.6 Sexe

L'intégration de cette variable est importante non seulement pour la prévision du phénomène de « Churn », mais aussi dans l'étape de prévention de celui-ci. Les preneurs de décisions doivent prendre en considération ces aspects dans leurs stratégies de commercialisions et doivent veiller à éviter toute discrimination.

Nous avons alors lié chaque identifiant avec son sexe correspondant sous la forme suivante :

	id	GENRE
0	87631	1
1	120851	1
2	28206	1
3	144072	1
4	78120	1
5	157648	1
6	97923	1
7	145007	0
8	125600	1
9	76149	1

FIGURE 3.8 – La base de données : Genre

3.3 Liaison des deux bases de données

Une fois que nos bases de données conséquentes sont prêtes, nous les fusionnons suivant l'identifiant du client (« id ») et la variable décrivant les dates de transactions (« DWP »). Nous obtenons une seule base de données contenant toutes les informations nécessaires pour notre modèle.

3.4 Nettoyage des données

Le nettoyage des données représente une partie importante dans la compréhension des données et dans l'assurance d'une qualité de prévision supérieure. Dans cette étape, nous nous chargeons de repérer et remanier ou ôter les observations qui souffrent d'erreurs de saisie ou de problèmes de redondance. Si jamais nous songions à éviter d'appliquer ce processus, nous aurions des estimations erronées avec des prédictions déplorables.

C'est à vrai dire l'origine des données qui est responsable de sa qualité, l'adaptation et le nettoyage de ces données est l'une des tâches d'un Data-Scientist. Nous connaissons plusieurs erreurs possibles, nous pouvons citer :

- Les erreurs de mesure
- Les erreurs de saisie de données
- Les possibilités de redondance

Le processus de nettoyage de données se fait comme suit :

3.4.1 Remplacement des valeurs manquantes

Après avoir recherché les valeurs manquantes se trouvant dans notre base de données, il s'est avéré qu'il en existe pour la variable « AGE » et la variable « pnb ». Nous pourrions tout simplement penser à supprimer les lignes qui souffrent de cette anomalie, mais ceci nous coûtera potentiellement des informations importantes pour l'entraînement du modèle.

DWP	0
id	0
AGE	11210
GENRE	0
pnb	130830
DIVORCED	0
MARRIED	0
OTHER	0
SINGLE	0
WIDOWED	0
AUTRE	0
CASABLANCA	0
FES	0
MARRAKECH	0
RABAT	0
SALE	0
TANGER	0
CREDIT_6_MONTH_AGO	0
CREDIT_5_MONTH_AGO	0
CREDIT_4_MONTH_AGO	0
CREDIT_3_MONTH_AGO	0
CREDIT_2_MONTH_AGO	0
CREDIT_1_MONTH_AGO	0
DEBIT_6_MONTH_AGO	0
DEBIT_5_MONTH_AGO	0
DEBIT_4_MONTH_AGO	0
DEBIT_3_MONTH_AGO	0
DEBIT_2_MONTH_AGO	0
DEBIT_1_MONTH_AGO	0
TARGET	0

FIGURE 3.9 – L'existance des valeurs manquantes

C'est pour cela qu'on remplace ces valeurs manquantes (« NULL Values ») dans la variable « pnb » par la valeur « 0 » et la variable « AGE » par sa valeur médiane.

DWP	0
id	0
AGE	0
GENRE	0
pnb	0
DIVORCED	0
MARRIED	0
OTHER	0
SINGLE	0
WIDOWED	0
AUTRE	0
CASABLANCA	0
FES	0
MARRAKECH	0
RABAT	0
SALE	0
TANGER	0
CREDIT_6_MONTH_AGO	0
CREDIT_5_MONTH_AGO	0
CREDIT_4_MONTH_AGO	0
CREDIT_3_MONTH_AGO	0
CREDIT_2_MONTH_AGO	0
CREDIT_1_MONTH_AGO	0
DEBIT_6_MONTH_AGO	0
DEBIT_5_MONTH_AGO	0
DEBIT_4_MONTH_AGO	0
DEBIT_3_MONTH_AGO	0
DEBIT_2_MONTH_AGO	0
DEBIT_1_MONTH_AGO	0
TARGET	0

FIGURE 3.10 – Enlever les valeurs manquantes

3.4.2 Élimination des doublons

Pour la plupart des bases de données résultantes, il s'avère qu'il y a plusieurs données redondantes, ils contiennent alors les mêmes informations. On se doit alors de les supprimer afin de ne pas donner des données inutiles au modèle.

3.5 Visualisation des données

3.5.1 Pourcentage des churners

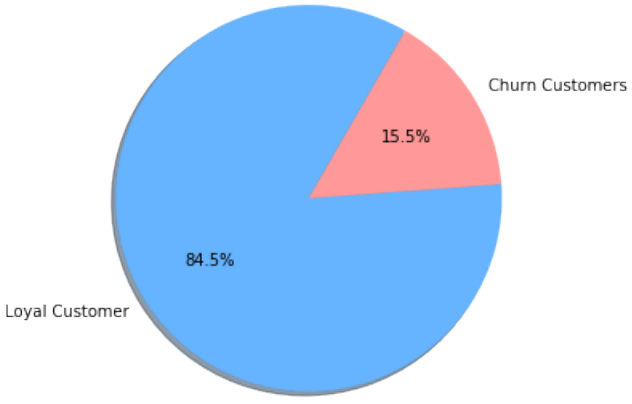

FIGURE 3.11 – Pourcentage des churners

Nous remarquons que 84.5% des clients sont toujours dans leur banque, alors que le reste a décidé de rompre sa souscription.

Au niveau de l'entrainement de l'algorithme du Machine Learning, ces résultats sont très satisfaisants.

3.5.2 Les clients selon leur sexe

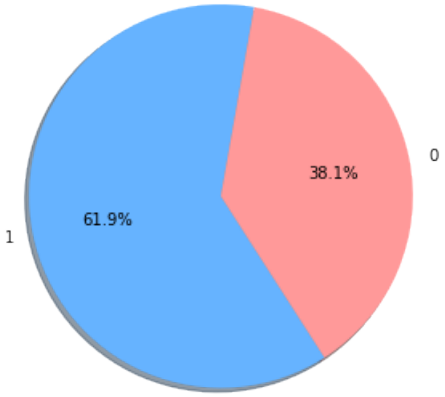

FIGURE 3.12 – Les clients selon leur sexe

Le sexe des clients joue évidemment un rôle très important dans la prévision du Churn. D'après les résultats obtenus, 38.1% représente la gente féminine alors que le reste représente les clients de sexe masculin.

3.5.3 La distribution des âges des clients

Nous remarquons que l'âge dominant dans notre base de données est aux alentours de 40. La figure ci-dessous nous montre la distribution de l'âge de nos clients.

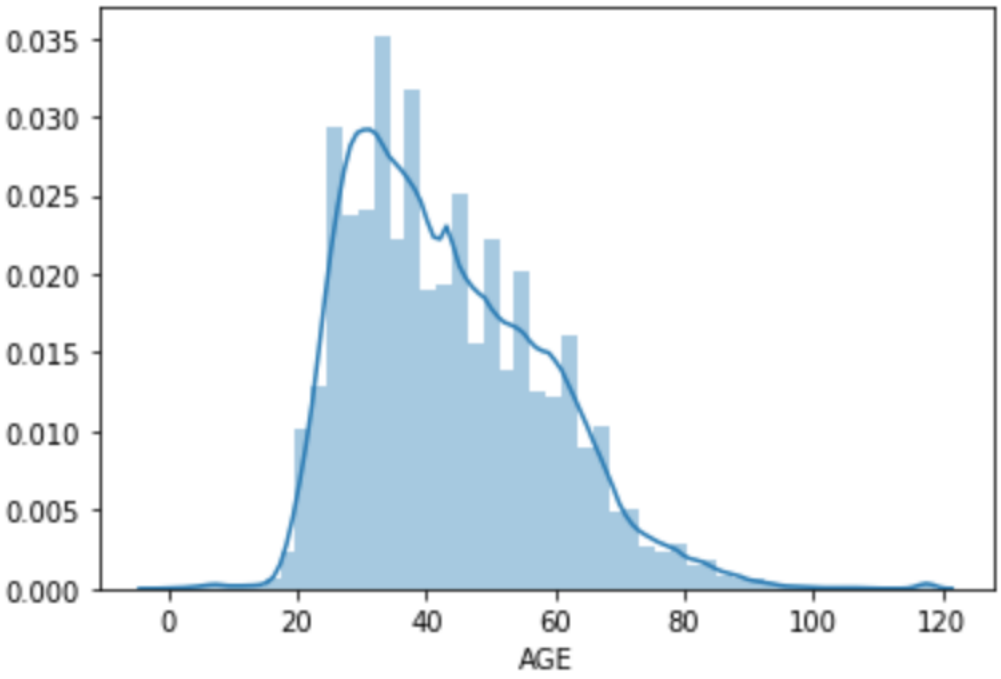

FIGURE 3.13 – La distribution des âges des clients

3.5.4 Répartition des churners dans les villes dominantes

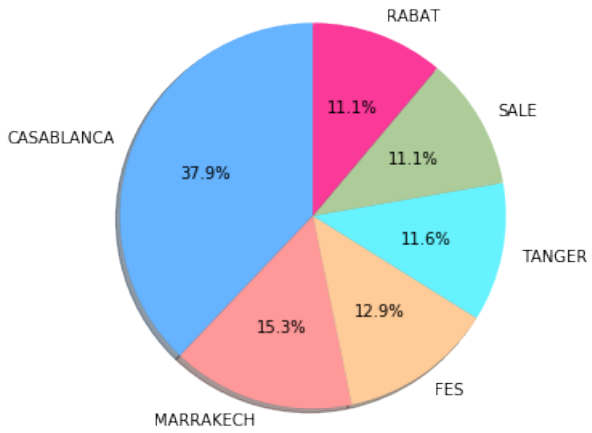

FIGURE 3.14 – Répartition des churners dans les villes dominantes

Chapitre 4

Résultat

4.1 Modélisation

Le processus de sélection est une étape importante dans la réalisation des modèles prédictifs. En effet, lorsque ce processus est accompli, on représente les données sous forme de matrice de telle façon à ce que les colonnes représentent les variables et les lignes qui décrivent les observations (ou individus). Ces variables sont ensuite fusionnées de façon à ce qu'elle forme une seule base de données de 30 colonnes qu'on utilise ensuite pour produire les algorithmes que nous utilisons pour construire les modèles. Les algorithmes de Data Mining sont utilisés pour construire ces modèles. Il s'agit d'algorithmes conçus à partir de données qu'on subdivise en deux sous-ensembles, l'ensemble d'apprentissage et de test. D'une part l'ensemble d'apprentissage est constitué de 70% des données et nous permet de construire l'algorithme, d'autre part, l'ensemble de test contient 30% des données et sert à mesurer la performance et la précision du modèle élaboré. Cet ensemble nous permet également de minimiser le surapprentissage et de régler les paramètres des algorithmes.

On parle de surapprentissage lorsque l'algorithme a enclin d'apprendre profondément les données qui lui permettent l'apprentissage. On dit qu'il existe un problème de généralisation. En d'autres termes, l'algorithme ne vas pas aboutir à des prédictions rigoureuses en ce qui concerne les résultats des nouvelles données encore non étiquetées. Durant la modélisation et pour chaque algorithme testé, on a recours à la classification. Ceci nous informe sur la nature de la variable cible, étant une variable de nature catégorielle.

4.2 Comparaison entre deux modèle choisis

4.2.1 Random Forest

	precision	recall	f1-score	support
0	0.93	0.96	0.95	309818
1	0.99	0.99	0.99	1699987
accuracy			0.98	2009805
macro avg	0.96	0.98	0.97	2009805
weighted avg	0.98	0.98	0.98	2009805

FIGURE 4.1 – Le résultat du modèle : Random Forest

4.2.2 La régression logistique

	precision	recall	f1-score	support
0	0.07	0.00	0.00	309818
1	0.85	1.00	0.92	1699987
accuracy			0.85	2009805
macro avg	0.46	0.50	0.46	2009805
weighted avg	0.73	0.85	0.77	2009805

FIGURE 4.2 – Le résultat du modèle : Régression logistique

Nous remarquons que la méthode du Random Forest est meilleure que celle de la régression logistique en termes de précision et de scoring.

4.3 Evaluation

4.3.1 La courbe ROC

Pour mesurer les performances d'un modèle de classification, la courbe ROC est une solution optimale pour avoir un résultat crédible à travers les seuils de classification. Avant de parler sur cette courbe, on doit bien définir les notions suivantes :

Quand le modèle prédit correctement la classe positive, on parle d'un résultat « **vrai positif** ». De façon analogue, quand le modèle prédit correctement la classe négative, on parle donc d'un résultat « **vrai négatif** » Or, un **faux positif** est un résultat où le modèle prédit incorrectement la classe positive. Et un **faux négatif** est un résultat où le modèle prédit incorrectement la classe négative.
La courbe ROC se base sur :

- Taux de vrais positifs : $TVP = \frac{VP}{VP + FN}$
- Taux de faux positifs : $TFP = \frac{FP}{FP + VN}$

Pour calculer les points d'une courbe ROC, il est plus efficace de calculer l'aire sous cette courbe, ou AUC, grâce à un algorithme de tri.

4.3.2 La courbe AUC

AUC signifie "aire sous la courbe ROC". Cette valeur mesure l'intégralité de l'aire à deux dimensions située sous l'ensemble de la courbe ROC.

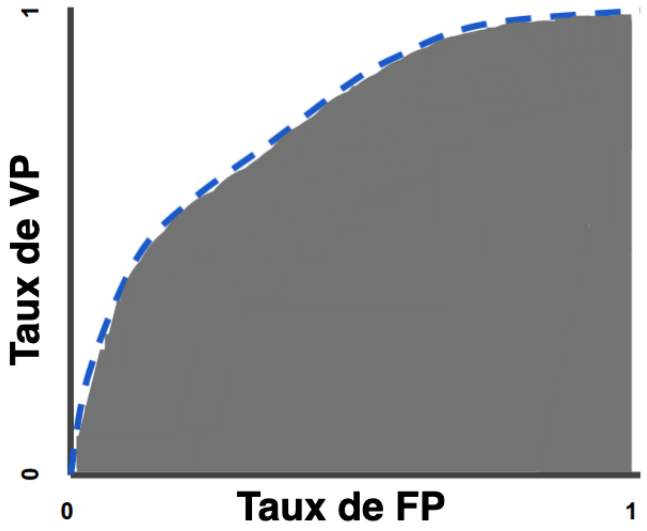

FIGURE 4.3 – Les courbes : ROC et AUC

Les valeurs d'AUC appartiennent à l'intervalle [0,1]. On a deux cas possibles :

- Si les prévisions sont erronées à 100%, le modèle a un AUC de (0,0)
- Si les prévisions sont correctes à 100%, le modèle a un AUC de (1,0)

Par exemple :

FIGURE 4.4 – Résultat du modèle de regression logistique

Parmi les avantages de L'AUC :

- L'AUC est invariante d'échelle : elle mesure la qualité du classement des prédictions, plutôt que leurs valeurs absolues.
- L'AUC est indépendante des seuils de classification : elle mesure la qualité des précisions du modèle quel que soit le seuil de classification sélectionné.

4.3.3 Application

On voit clairement la puissance du modèle "Random Forest", d'un coefficient de $\boxed{AUC = 0.975 = 97.5\%}$

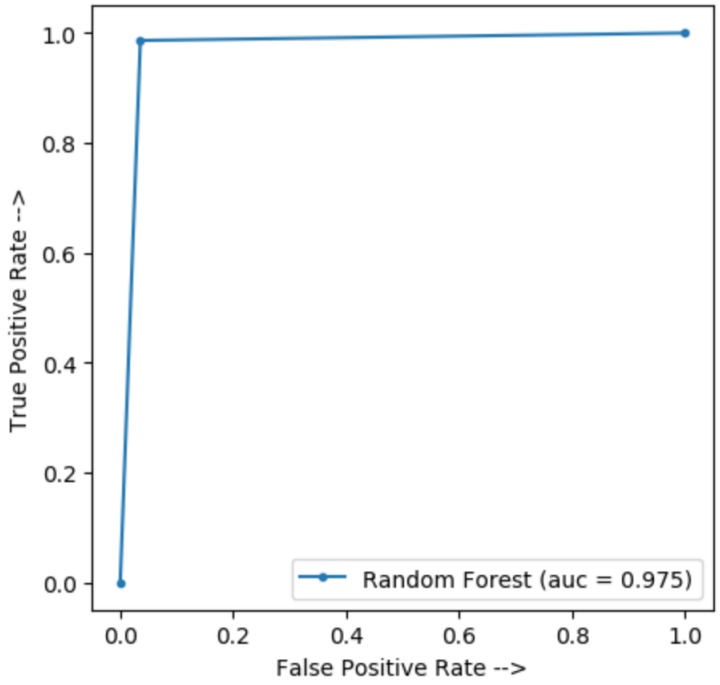

FIGURE 4.5 – Courbe AUC du modele : Random Forest

La modèle de la régression loqistique a : $\boxed{AUC = 0.807 = 80.7\%}$

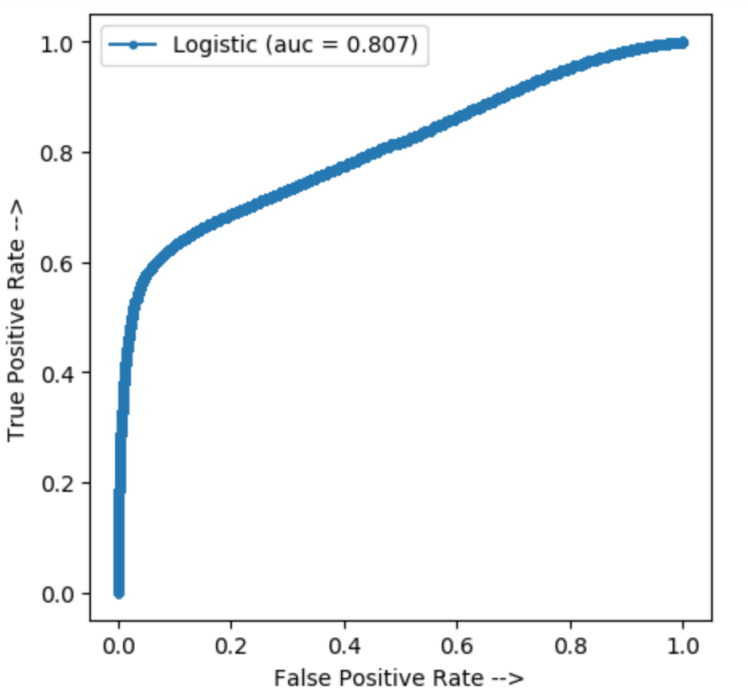

FIGURE 4.6 – Courbe AUC du modele : Régression Logistique

4.3.4 Comparaison

Nom	Type de problème	Interprétation	Précision prédictive	Vitesse d'apprentissage	Fonctionne avec un petit nombre d'observation	Manipulation des variables non-pertinentes
Régression logistique	Classification	Moyenne	Faible	Rapide	Oui	Non
Random Forests	Classification & régression	Faible	Haute	Lente	Non	Oui

TABLE 4.1 – Etude comparative des deux algorithmes de prédiction

Conclusion

Le problème est basé sur le domaine du secteur bancaire où la banque veut prédire le taux d'attrition d'un client en fonction des données précédentes de ce dernier. Par "churn", on entend que la banque veut prédire si un client sera en défaut de paiement au cours du prochain trimestre en fonction de ses antécédents de crédit. Le principal problème est de prévoir si un client sera en défaut de paiement ou non en fonction des données antérieures du client. Du point de vue d'une banque, il est important de maintenir les relations d'affaires et les relations avec les clients/ En outre, si l'on peut prévoir qu'une personne sera défaillante, des mesures primitives peuvent être prises afin de s'assurer que de telles violations ne se produisent pas. Toute entreprise doit comprendre la perte de clientèle et les coûts importants qui en sont engendrés. Ils sont parfaitement conscients qu'ils ont besoin d'un moyen d'établir une plus grande confiance avec leurs clients et de créer une véritable loyauté. Le processus du traitement du Churn est par conséquent nécessaire pour le fonctionnement sain d'une entité financière quelconque. En fait, des actions de vente et de marketing efficaces sont censées influencer le comportement des clients, avec pour conséquence un changement du modèle lui-même. L'analyse des changements de règles et du classement des attributs représente une occasion unique de mesurer le succès réel des stratégies commerciales de la banque.

Bibliographie

1. What is the Purpose of Data Science ? Know Its Importance. DataFlair
   https ://data-flair.training/blogs/purpose-of-data-science/
2. What is Big Data why is Big Data important in todayís era
   https ://medium.com/@syedjunaid.h47/what-is-big-data-why-is-big-data-important-in-todays-era-8dbc9314fb0a
3. What Is Machine Learning ? NetApp
   https ://www.netapp.com/artificial-intelligence/what-is-machine-learning/
4. Why Do Customers Stop Doing Business With a Bank ? Noah Mukhtar
   https ://medium.com/@noah.fintech/creating-a-banking-customer-churn-model-1a2d0850f071
5. When Should I Use Regression Analysis ? Statistcs By Jim
   https ://statisticsbyjim.com/regression/when-use-regression-analysis/
6. Classification - Machine Learning. SimleLearn
   https ://www.simplilearn.com/classification-machine-learning-tutorial
7. Classification : Basic Concepts, Decision Trees, and Model. Kumar
8. Logistic Regression ó Detailed Overview. Saishruthi Swaminathan
   https ://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
9. Random Forests Algorithm. Michael Walker
   https ://www.datasciencecentral.com/profiles/blogs/random-forests-algorithm
10. DECODING THE POPULARITY OF JUPYTER AMONG DATA SCIENTISTS
    https ://www.analyticsinsight.net/decoding-popularity-jupyter-among-data-scientists/
11. Why Python Programming Language is important in Data Science ? Aaksh Kumar
    https ://medium.com/javarevisited/why-python-programming-language-is-important-in-data-science-beb4a7f91f75