RAPPORT DE STAGE TECHNIQUE

MACHINE LEARNING

Etudiant : RAMI LETAIEF
Majeure : Business Intelligence

Tuteur de stage : Atef KOURAICHI

Réfèrent pédagogique : Hassane MIMOUN

TABLE DES MATIERES

REMERCIEMENTS

Je tiens à adresser mes profondes appréciations et gratitudes à tous ceux qui m'ont aidé de près ou de loin à réaliser ce modeste projet.

Une mention spéciale va à l'équipe enthousiaste de Orange Tunisie pour m'avoir rendu la période de stage agréable. Je profite de cette occasion pour exprimer ma reconnaissance à M. Mounir MELLITI, chef de département Big Data et BI à Orange Tunisie, et M. Atef KOURAICHI pour leur rôle fondamental dans mon stage de M1. Ils m'ont fourni tous les conseils, l'assistance et l'expertise dont j'avais besoin au cours de mon expérience professionnel.

Je suis profondément reconnaissante à tous mes professeurs qui ont contribué à ma formation durant mon parcours universitaire et qui m'ont beaucoup conseillé pour le choix du stage et du projet professionnel.

Je voudrais remercier ma chère et affectueuse famille pour son soutien et ses sacrifices, ils m'ont beaucoup aidé à trouver le stage qui m'intéressait.

Ce dernier mot de reconnaissance que j'ai gardé va aux membres du jury pour le temps et les efforts qu'ils ont consacrés à l'évaluation de mon travail.

INTRODUCTION

Dans le cadre de ma formation d'ingénieur à EFREI paris, j'ai réalisé mon stage technique de 4e année chez orange Tunisie.

Le but principal de ce stage a été de me faire découvrir un environnement professionnel et de m'aider à développer mes compétences.

J'ai alors décidé de faire mon stage à l'étranger, plus précisément en Tunisie qui est mon pays natal pour découvrir le monde du travail local.

J'ai réussi à obtenir un stage dans une spécialité qui m'intéressait beaucoup, qui est le machine Learning. Mon stage s'étend sur une durée de 5 mois, et j'avais pour sujet « le machine Learning pour le Market Share régional ».

J'ai donc pu intégrer le service DSI et découvrir beaucoup de choses durant ces 5 mois. J'ai eu l'occasion d'acquérir de nouvelles compétences, découvrir de nouveaux domaines qui étaient lies à ma formation, mais que l'on n'avait pas encore aborde.

Durant ce stage j'ai pu travailler sur 2 sujets différents mais qui portaient sur le même thème qui est « le Market Share régional ». Ces 2 projets de machine Learning étaient très intéressants car j'ai pu travailler a deux sur le premier et seul sur le deuxième.

Après une présentation de la société orange, je détaillerai mes 2 missions ainsi que le travail réalisé dans leur cadre.

PARTIE 1 : ORANGE, TOUJOURS CONNECTE, JAMAIS SANS MOBILE, JAMAIS SANS INTERNET

I. ENTREPRISE D'ACCUEIL

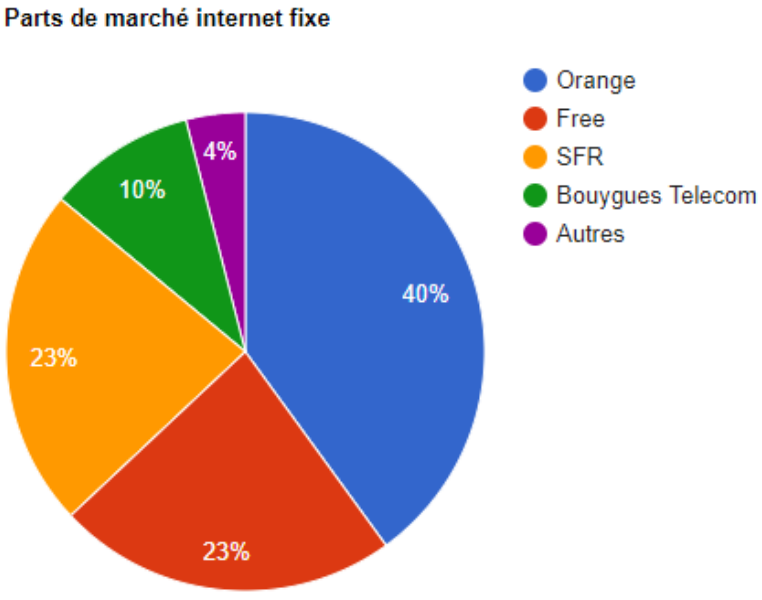
Dans cette partie, je commencerai par la présentation du groupe orange base en France puis je passerai à l'organisme d'accueil à savoir orange Tunisie dans lequel j'ai pu effectuer mon stage de fin d'études et plus précisément dans le département bi et big data qui m'a accueilli pendant mes cinq mois de stage.

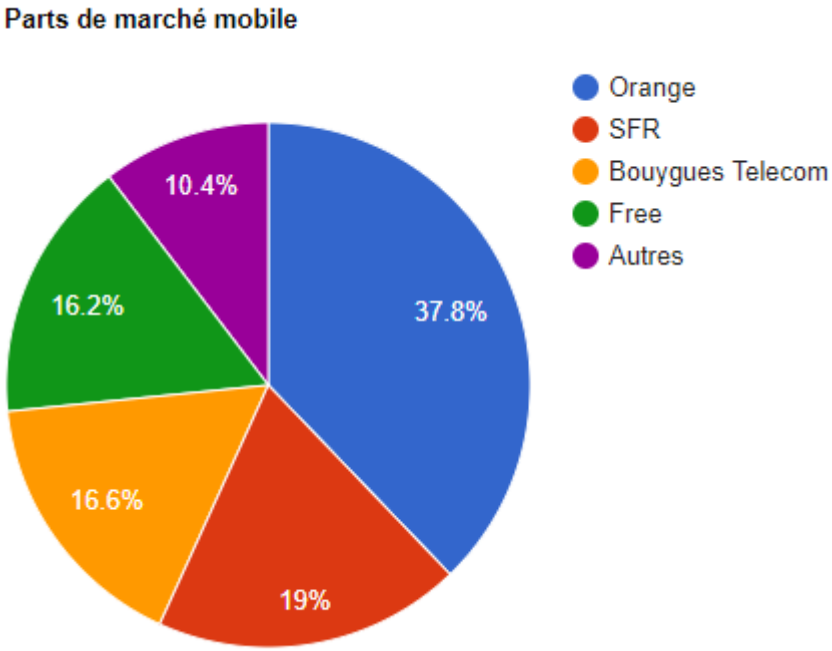
a) ORANGE DANS LE MONDE

Orange est actuellement une entreprise française de télécommunications. Elle est aussi la marque phare de France Telecom qui est l'un des principaux opérateurs de télécommunications dans le monde.

France Telecom, qui est un exploitant autonome de droit public en 1991, passe de statut d'exploitant public à société anonyme en 1996 (avec comme seul actionnaire l'état français). Par la suite, elle a ouvert son capital et devient cotée en bourse (Paris et New York). En 2000, elle fit l'acquisition de l'opérateur Britannique Orange Plc pour 40 millions d'euros. En septembre 2004, l'état français cède une partie de ses actions et France Telecom devient alors une entreprise privée. À partir de 2006, la plupart des activités du groupe passent sous la marque et le logo orange.

Orange est le leader du marché français des opérateurs téléphoniques. En 2015, elle comptait 262.9 millions de clients dans le monde. En 2013, l'entreprise est leader ou second operateur dans 75 % des pays européens ou elle est implantée et dans 83 % des pays en Afrique et au Moyen-Orient, dont la Tunisie.





b) ORANGE TUNISIE : LE TRAVAIL ENSOLEILLE

Orange Tunisie est le deuxième opérateur prive de télécommunications en Tunisie depuis mai 2010. Cependant, il est le premier à obtenir une licence pour exploiter le réseau 3g en Tunisie. Il est le fruit d'une alliance entre Orange Telecom et la société Investec. 2 mois après son lancement, il a pu se différencier de ses concurrents et a réussi à attirer 650 000 abonnes. Par ailleurs, en 2014 le nombre d'abonnés a dépassé les deux millions.

Pour répondre aux attentes de ses clients, Orange Tunisie s'appuie sur un grand marché de vente directe et indirecte qui dépasse 400 points de vente. Ce réseau changera selon la couverture géographique d'orange Tunisie.

L'opérateur a toujours été un bâtisseur dans le marché de télécommunications, en proposant de nouveaux services avec des technologiques avancées, notamment la technologie 4g dont la couverture réseau couvre 60 % de la population tunisienne.

Grace a son innovation technologique et son savoir-faire, orange Tunisie a réussi à attirer plus de 3 363 299 abonnés mobiles et à devenir l'opérateur préfère sur le marché tunisien, en essayant d'améliorer ses services et fidéliser ses clients.

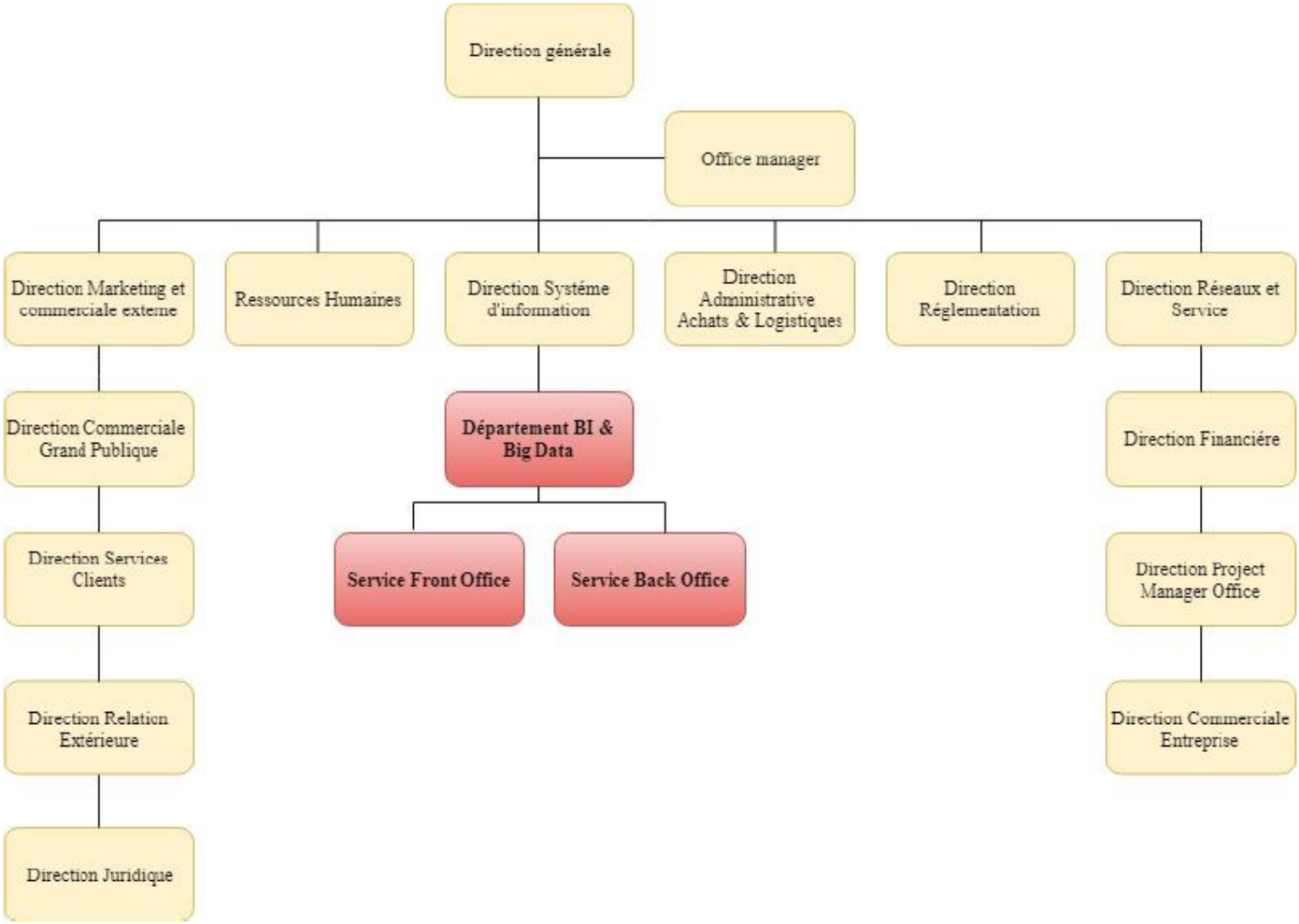
c) L'ORGANISME D'ACCUEIL

L'entreprise englobe quatorze directions avec différents domaines d'activités qui coopèrent dans le but de garantir une meilleure qualité de service à ses clients.

Pendant la période de mon stage, j'ai effectué ma mission au sein du département big data et bi, dont les principales missions sont :

- La mise en place des systèmes informatiques décisionnels dont le but d'accélérer et d'améliorer la prise de la décision.
- Le recueil, l'analyse des besoins et l'optimisation des processus métiers internes.
- La mise en place d'une base de données intermédiaire (ods) pour la collecte, le nettoyage et l'intégration des données à partir de sources différentes.
- La mise en place d'un entrepôt de données (DWH) et d'un magasin de données (datamart) pour le stockage des données d'achats, de ventes ou de stocks…
- La mise en place d'une solution ETL pour le chargement et le traitement des données entre les systèmes sources, l'ODS, le DWH ou les datamarts.
- L'augmentation de l'efficacité opérationnelle.
- La génération de nouvelles recettes.
- Création de programmes de machine Learning

Il est indéniable que cette société doit sa réputation de leader aussi bien à son organisation qu'a sa hiérarchie bien étudiée. Pour mieux comprendre la disposition des différents départements au sein de l'entreprise la figure suivante illustre la hiérarchie d'orange Tunisie.



ORGANIGRAMME : L'ORGANISME D'ACCUEIL

Mon stage a eu lieu au sein de la direction système d'information (DSI), spécifiquement dans le département BI et Big Data.

En fait, la DSI accompagne les directions métiers dans l'adoption des solutions décisionnelles afin de suivre l'activité de l'entreprise. Pour ce faire, orange Tunisie a mis en place une plate-forme et des solutions de collecte et de restitution de données qui peuvent fournir une meilleure prise de décision par les responsables. La DSI représente un maillon stratégique dans la chaine décisionnelle. En effet, chez orange Tunisie, la DSI permet de fournir un aperçu global sur les activités de l'entreprise, anticiper les risques et répondre aux besoins émergents.

Pour ma part lors de ce projet, j'ai été inclus dans le service back office et l'on travaillait en parallèle avec le service front office pour certains besoins. Mon équipe était composée de 6 ingénieurs.

d) ORANGE TUNISIE A VOTRE SERVICE

Orange Tunisie est une société innovante et compétitive, elle veille toujours à développer et à commercialiser des services montrant qu'elle n'est pas un simple opérateur de télécommunication, mais bien un intégrateur de solutions. Parmi les services fournis nous pouvons citer :

- Les services de communications résidentiels (SCR) se traduisant par la téléphonie fixe, l'internet a bas débit (par modem), haut débit (par ADSL) ou très haut débit (par la fibre optique). La téléphonie IP, la visiophonie, la télévision numérique et les contenus multimédias. De plus orange Tunisie fournit un appareil électronique flybox. C'est un boitier qui fait office de modem routeur wifi. Cet appareil permet de partager la connexion a plusieurs types de terminaux, le client peut donc bénéficier d'un accès a plusieurs types de services sans être oblige d'utiliser les câbles.

- Les services de communications personnels (SCP) appelés aussi services mobiles. Ainsi, orange Tunisie offre à ses clients des puces électroniques compatibles avec les appareils téléphoniques. Une fois que la puce est activée, le client peut bénéficier d'un accès au réseau mobile.

- Les services de communications d'entreprise (SCE).

PARTIE 2 : APPRENDRE L'APPRENTISSAGE AUTOMATIQUE, IRONIQUE ?

I. RAPPEL DU SUJET ET OBJECTIFS PRINCIPAUX

Durant mon stage chez orange, le sujet était le market share regional et les objectifs principaux étaient variés mais aussi lies car ils visaient tous à impacter l'étude du market share avec différentes techniques et outils (comme le machine Learning ou le travail sur les data Warehouse de la boite).

Mon tuteur de stage, qui est aussi le chef de service back, m'a permis de pouvoir travailler avec beaucoup de liberté, ce qui m'a beaucoup aidé à me développer et progresser rapidement.

Je n'ai pu réellement comprendre ma mission qu'en arrivant au stage car je ne connaissais pas encore réellement les attentes de mes supérieurs. Les taches variaient beaucoup d'un jour à l'autre mais le but final était la prédiction de certaines informations pour les aider lors de leurs décisions. C'est aussi à ce moment que j'ai su que j'allais travailler avec un collègue sur une première mission qui est le churn des abonnes. Pour ensuite enchainer sur un autre projet solo qui est la prédiction de dépenses des clients avec des abonnements prépayés.

a) LE CHURN DANS TOUS SES ETATS

Le terme churn désigne la proportion de clients contractuels qui quittent un fournisseur au cours d'une période donnée. Il existe deux principaux types de désabonnements, à savoir le désabonnement volontaire et le désabonnement involontaire. Il y a désabonnement volontaire lorsque le client a initié la résiliation du service. A contrario le churn involontaire signifie que la société a suspendu le service du client, ce qui est généralement du a un non-paiement ou à un abus de service.

Ce phénomène est très répandu sur des marches très concurrentiels tels que le secteur des télécommunications. Ainsi plusieurs entreprises qui étaient établies sur cette marche et détenaient un monopole, voient maintenant celui-ci disparaitre au profit d'un oligopole ou même d'un marché concurrentiel. Cela signifie que les produits et surtout les services offerts par ces compagnies ne sont plus les seuls sur le marché.

Partant de ce constat, il est impératif de distinguer les clients qui ne sont pas réticents à se tourner vers un concurrent. Par conséquent, la prévision du taux de désabonnement des clients est devenue un problème essentiel dans le secteur des télécommunications.

Dans le but d'encourager les clients à consommer davantage, les opérateurs téléphoniques doivent mettre en œuvre des stratégies de fidélisation clients. Celles-ci devront déterminer les caractéristiques, valeurs, comportements, attitudes de leurs clients et tous autres indices, qui permettront l'amélioration du processus de rétention, d'où provient le besoin pour ce projet.

Le défi est de détecter les comportements de consommation client en termes de valeur et de fréquence. Ceci permettra de pouvoir cibler les clients qui ont de fortes probabilités de suspendre leur service.

L'apprentissage automatique (machine Learning) peut prendre le relais en se basant sur l'analyse comportementale de la clientèle. Il permet de proposer une expérience personnalisée aux clients cibles et de les sensibiliser via les campagnes marketing. L'objectif est alors de détecter les clients, à forte valeur ajoutée ou à faible usage afin de les fidéliser.

b) LA POPULATION CIBLE

Afin de cibler la population d'abonnes susceptible d'abandonner le service, orange Tunisie mise à mettre en place un modèle prédictif permettant de prédire lesquels de ces derniers vont partir et ceci en passant par l'analyse comportementale de ces clients dans le but d'anticiper leur décision de quitter le service.

Dans le cadre de ce projet, nous contribuons aux taches de data scientist allant de la compréhension du domaine des télécommunications, la collecte et analyse des données à l'implémentation des modèles prédictifs de classification. Plus précisément, afin de construire des modèles d'apprentissage automatique pour la prédiction de churn, il nous est d'abord demande d'accomplir deux tâches principales :

1. Identifier et segmenter la population cible.
2. Identifier les variables indicatrices et influentes (kpi's).

Le résultat de ces modèles prédictifs est nécessaire pour fournir aux développeurs bi de orange Tunisie des informations pertinentes et les accompagner dans leurs prises de décisions dans le but de minimiser le taux de désabonnement.

c) QUI VEUT DES RECHARGES ?

Le deuxième projet sur lequel j'ai pu travailler était aussi un projet de machine Learning. Ce projet que j'ai pu réaliser seul grâce aux connaissances et techniques acquises durant le premier projet. Son but était de prédire les recharges des clients ayant un abonnement prépayé. Il faut donc savoir qu'une grande partie de la population en Tunisie possède des abonnements prépayés, c'est donc la principale source de revenus de la compagnie.

Ici aussi, il faudra procéder de la même façon que pour le projet de churn, c'est-à-dire toute l'étude se base sur le comportement des utilisateurs. Le choix des kpis est similaire aussi.

II. ENVIRONNEMENT TECHNIQUE

Avant d'entrer dans les détails de notre projet, il faut d'abord procéder par une étude préliminaire de quelques concepts. Nous consacrons cette partie à expliquer les concepts utiles et les connaissances théoriques facilitant la compréhension du travail à venir.

a) APPRENTISSAGE AUTOMATIQUE

L'apprentissage machine est une discipline qui s'appuie sur des outils mathématiques et statistiques en vue de l'analyse et prévision des modèles dans les données. Le diagramme de la figure qui suit illustre le cycle de vie typique pour la création d'un modèle d'apprentissage

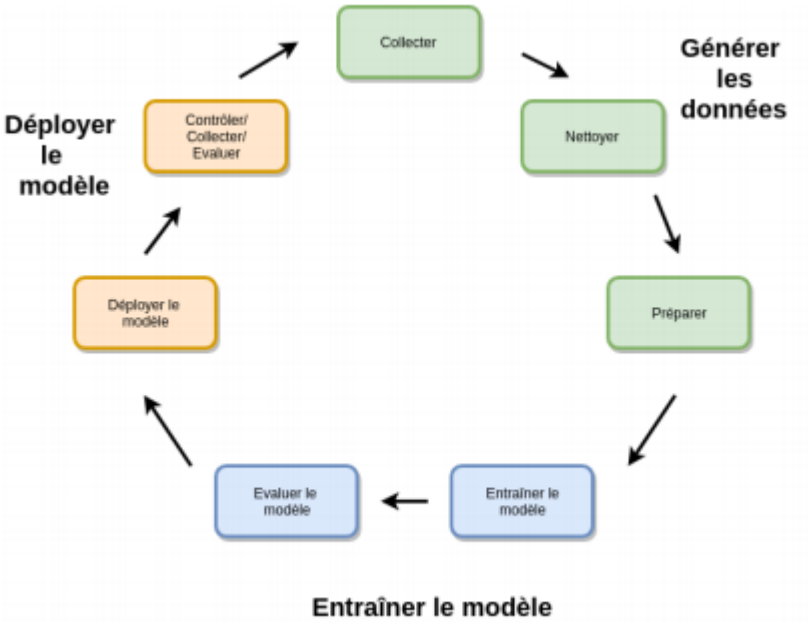


FIGURE : PIPELINE DE L'APPRENTISSAGE AUTOMATIQUE

Nous "enseignons" à un ordinateur pour faire des prédictions ou des inférences. Tout d'abord, nous utilisons un algorithme et des exemples de données pour former un modèle. Ensuite, nous intégrons le modèle à notre application pour générer des inférences en temps réel et à grande échelle. Dans un environnement de production, un modèle apprend généralement des millions de données et produit des prédictions a moins de 20 millisecondes.

Nous distinguons deux grandes familles de l'apprentissage machine :

1) Apprentissage supervise. Les algorithmes d'apprentissage supervise tentent de modéliser les relations et les dépendances entre la sortie et les entités en entrée, de manière à

pouvoir prédire les valeurs de sortie des nouvelles données en fonction des relations apprises à partir de l'ensemble de données précédentes.

En d'autres termes, nous recevons un échantillon de données labellisées ((x1, y1), (x2, y2), (x3, y3), ...) Et nous visons à prédire le résultat des nouvelles observations (x [i] → y [i]).

Les problèmes à résoudre avec un apprentissage supervise sont du type : détection des activités frauduleuses dans les transactions bancaires, classification des emails en tant que spam et non spam ou alors prévision de la demande des produits.

2) Apprentissage non supervise. Dans cette catégorie d'apprentissage il n'y a pas de catégories ou d'étiquettes de sortie sur lesquelles l'algorithme peut essayer de modéliser des relations. Ils sont principalement utilisés dans la détection de modèles (pattern en anglais) et la modélisation descriptive.

En d'autres termes, nous ne recevons que des observations brutes de variables aléatoires (x1, x2, x3, ...) Et nous espérons découvrir la relation avec les variables latentes structurelles (x [i] → y [i]).

Ces algorithmes essaient d'utiliser des techniques sur les données d'entrées pour rechercher des règles, détecter des modèles, et résumer et regrouper les points de données permettant de générer des informations significatives et de mieux décrire les données aux utilisateurs.

Les cas d'utilisation sont principalement le regroupement (cluster en anglais), la détection des anomalies et la réduction dimensionnelle.

Les deux familles ont leurs propres techniques. Par exemple, nous pouvons distinguer :

- Les problèmes de régression et problèmes de classification qui peuvent être résolus via un apprentissage supervise.

1. Les problèmes de régression prédisent les valeurs prises par une variable dépendante quantitative, base sur des variables explicatives quantitatives et/ou qualitatives (caractéristiques).

2. Les problèmes de classification prédisent l'appartenance des objets (observations, individus) a une classe (ou catégorie) basée sur des facteurs quantitatifs et/ou variables explicatives qualitatives (appelées caractéristiques).

- Les problèmes de regroupement entrent dans l'apprentissage non supervise.

b) LES KPIS

Afin de permettre à l'opérateur orange Tunisie de faire le suivi de ses clientèles et d'obtenir en permanence des informations sur leur consommation, habitudes d'achat et leurs

réclamations par rapport à la qualité des services offertes par l'opérateur, nous avons défini des indicateurs de performance appelés kpi's (key performance indicators) permettant de spécifier le comportement de ses abonnes.

En effet, un indicateur de performance kpi (indicateur de performance clé) est une mesure ou un ensemble de mesures centre sur un aspect critique de la performance globale de l'organisation qui contribue à l'évaluation de la situation par un décideur.

Les indicateurs doivent être soigneusement choisis en accord avec les besoins spécifiques et la stratégie poursuivie. Pour choisir correctement nos indicateurs, nous avons suivi des critères de choix :

- ✓ Associe à un ou plusieurs objectifs précis : les indicateurs choisis doivent mesurer la performance selon les objectifs choisis.
- ✓ Induit toujours à une décision : un indicateur clé entraine toujours les décisions nécessaires.
- ✓ Utilisable en temps réel : pour piloter, il faut disposer de l'information réactualisée lorsque la décision est nécessaire.
- ✓ Être mesurable : un bon indicateur doit être construit facilement, ne nécessite pas des données inaccessibles et facile à comprendre.

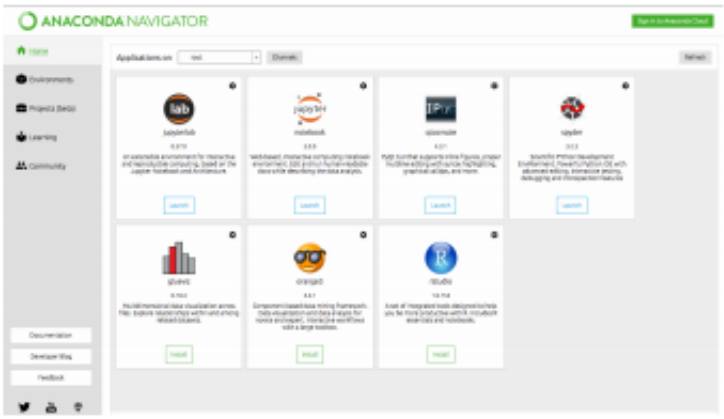
c) ENVIRONNEMENT LOGICIEL

Notre travail a été mis en œuvre au moyen des outils logiciels suivants :

- ❖ TOAD

Toad est un logiciel développe par la société quest software permettant la consultation et l'administration d'une base de données. Ce logiciel est, en particulier, utilisé par les développeurs oracle ainsi que les administrateurs de bases de données.

❖ ANACONDA



La distribution open source d'anaconda est le moyen le plus simple d'exercer la science des données avec les langages python / r et l'apprentissage automatique sur différents systèmes d'exploitation : linux, Windows et mac os x.

d) TECHNOLOGIES UTILISEES

❖ PYTHON 3

Python est un langage de programmation puissant et facile à apprendre. Il possède des structures de données solides et accorde une approche propre et efficace à la programmation orientée objet. La syntaxe du langage de programmation python étant conçue pour être élégante et facilement lisible, elle convient parfaitement au script et au développement rapide d'applications dans de nombreux domaines, tels que l'exploration de données et l'apprentissage automatique, car elle offre un nombre illimité de bibliothèques implémentant des modèles et des outils d'apprentissage automatique (sklearn, spacy, nltk, etc.).

❖ LIBRAIRIES

✓ SCIKIT-LEARN

Scikit-learn est un kit permettant de développer des applications d'apprentissage automatique, mettant en œuvre les méthodes de classification, de régression (prédiction) ou groupement avec python. La solution est open source et provient de l'Inria (institut français de recherche en informatique et en automatisme) et de Telecom paris tech. Sa communauté internationale compte 1135 contributeurs depuis sa création avec environ 70 actifs par version.

- ✓ MATPLOTLIB

Matplotlib est une bibliothèque du langage de programmation python permettant de dessiner et de visualiser des données sous forme de graphiques.

- ✓ GRAPHVIZ

Graphviz (abrège de graph visualization software) est un ensemble de fonctionnalités open source crées par les laboratoires de recherche d'att. Il permet la manipulation des graphes définis à l'aide de scripts suivant le langage dot.

- ✓ PANDAS

Pandas est une bibliothèque du langage de programmation python très populaire en matière de manipulation et d'analyse des données. Elle offre beaucoup de fonctionnalités, entre autres, des structures de données et des opérations de manipulation de séries temporelles et de tableaux numériques.

PARTIE 3 : DE L'APPRENTISSAGE A LA REALISATION

Dans cette section, nous allons dresser le travail réalisé sur les deux projets de machine Learning au cours de nos cinq mois de stage acharne.

I. ETUDE SUR LES DONNEES

a) COMPREHENSION DU DOMAINE

Face à une concurrence ardue, les opérateurs téléphoniques voient leurs services en compétitivité sur le marché. En outre, les exigences de la clientèle deviennent de plus en plus compliquées et difficile à satisfaire. Ceci mène certaines entreprises à régler leur relations clients pour éviter le risque de perte ou attrition.

Pour éviter ce fléau, nous avons identifié les besoins de orange Tunisie qui sont l'identification de la population et l'implémentation d'un modèle prédictif pour prédire ceux qui ont une forte probabilité de suspendre le service. Ceci dans le but de diminuer le risque de l'attrition.

b) COLLECTE DES DONNEES

Les opérateurs téléphoniques sont perçus comme une quantité gigantesque de données stockant les informations des trafics des abonnées à travers un réseau.

Chez orange Tunisie, les données dont nous disposons proviennent essentiellement des deux sondes : astellia et otarie.

✧ ASTELLIA

La sonde astellia est l'une des grandes sources des données des clients d'orange Tunisie. En effet, elle permet de capturer les évènements des usages voix et messagerie effectues entre le client et le réseau en offrant une optimisation automatisée, des aperçus géolocalises et des analyses de données volumineuses.

✧ OTARIE

Implantée dans les réseaux d'accès, la sonde otarie mobile web est un outil de suivi des usages data et du détail du trafic web. A l'aide des techniques de décryptage et d'interprétation des flux générés la sonde permet de capter et de suivre les informations collectées par une activité data et le détail du trafic web pour les clients résidentiels (ADSL) et mobile

Ces données peuvent prendre donc différentes formes : voix, image, données de navigation etc. et elles sont de différentes sources. Ces dernières sont les bases de données relationnelles (ERP, CRM, BSCS), les sondes otarie (connexion data) et astellia (2 types : 2g et 3g) et les fichiers générés à partir de la médiation.

Ces derniers sont une suite de CDRS. Chaque ligne du fichier constitue un CDR qui maintient les informations relatives à chaque type de trafic (voix, sms, données) ainsi que d'autres détails utiles sur chaque appel (les numéros des parties de la communication, date et heure d'appel, durée de l'appel, identificateur du central téléphonique). Pour la préparation des données, nous avons interrogé essentiellement les sondes otarie et astellia, les figures en dessous représentent respectivement les tables 'astellia_gsm_filtered', 'astellia_iu2', 'otarie_appsession_agg'.

Champs	Définitions
Con_duration	Durée de la connexion en ms
mobile_type	Le type de téléphone
insert_date	La date de la connexion
ci	L'identifiant de la cellule
Start_time	Temps de début de la connexion
End_time	Temps de fin de la connexion

FIGURE : TABLE ASTELLIA_GSM_FILTERED

Champs	Définitions
Start_time	Temps de début de la connexion
End_time	Temps de fin de la connexion
Con_duration	Durée de la connexion en ms
imsi	International mobile subscriber id du A number permet d'identifier un usager
imei	International mobile equipement identity de l'abonné
lac	Le code de la zone de localisation actuelle
sac	Ville /cellule du début de connexion
Service_type	Type du service : voix,sms...
Ringing_duration_ms	La durée de sonnerie en connexion voix

FIGURE : TABLE ASTELLIA_IU2

Champs	Définitions
Con_nb	Identifiant associé à chaque évènement.
Start_time	Temps de début de connexion.
Duration	La durée associée à chaque connexion.
Imei	International Mobile Equipement Identity de l'abonnée.
Msisdn	Le numéro de l'abonnée.
App	Identifiant pour la catégorie d'application 1. Web : données provenant du navigateur Web 2. Mail : échanges de courrier électronique 3. Streaming : échange de données vidéo ou audio en streaming
Loc_info	Masque binaire hexadécimal (16 caractères) pour l'extrait d'informations sur l'emplacement de l'utilisateur.

FIGURE : TABLE OTARIE_APPSESSION_AGG

II. PREPARATION DES DONNEES

Nous attaquons l'une des phases les plus délicates et les plus couteuses en termes de temps comme nous l'avons mentionné dans la phase préparation des données chapitre précèdent : l'identification des indicateurs clés de performance.

Dans cette partie nous exposons le choix de la population choisie, la sélection des données, la fusion et croisement des tables existantes ainsi que les variables calculées.

a) CHOIX DE LA POPULATION

Pour la prédiction des abonnes susceptibles d'un éventuel départ, nous avons bases notre étude sur la population dont le contrat est du type 'prépayé'. Nous avons travaillé sur le dernier appel émis par ceux-là. Cette donnée se trouve dans les bases de données concernant les appels entrants et sortants de orange Tunisie. Cette base dans le langage de télécommunication s'appelle table 'last usage'. Pour le deuxième projet qui est la prédiction des recharges en fonction du comportement, j'ai eu comme tache de rajouter deux colonnes et modifier le job Talend qui possède le code SQL de récupération de données de la table last_usage. Le choix de la population est similaire pour les deux projets, seul la variable flag change.

Ainsi les abonnes n'ayant pas émis d'appels pendant trois mois nous les avons étiquetés comme étant une population churner flag '1', et ceux qui ont émis un appel sortant pendant cette durée ont été étiquetés non churner flag '0'. Soit un total de 10000 clients dont 9274 clients en service contre 726 clients ont quitté le service. Pour le second projet, le flag sera le montant de recharge de la période sur laquelle on va étudier la population.

L'extraction de cette population a été faite avec des requêtes SQL sur les tables de la DWH (data Warehouse) ou sont stockes les flux des données générés par les deux sondes (voir annexe script SQL pour la requête d'extraction). La figure ci-dessous nous montre la proportion des clients en pourcentage.

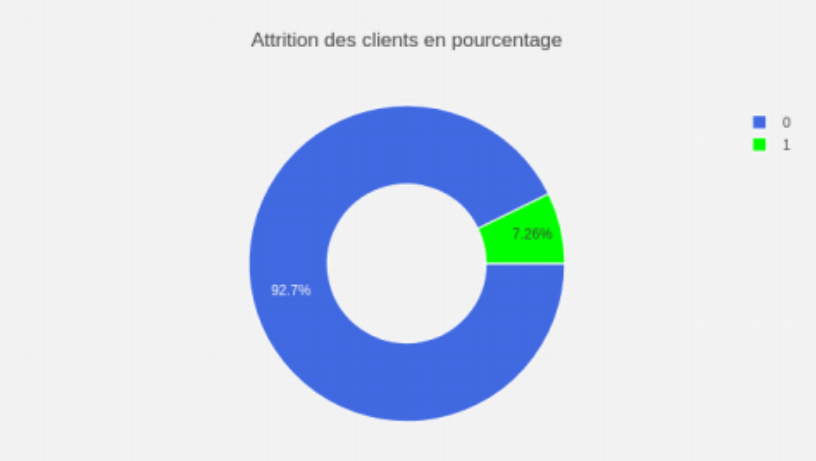


FIGURE : PART DES CLIENTS EN POURCENTAGES

• le pourcentage en bleu représente les abonnes qui sont encore actifs

• le pourcentage en vert représente les abonnes qui ont quitté le service.

b) SELECTION ET EXTRACTION DES DONNEES

La sélection des caractéristiques est l'étape qui consiste à sélectionner minutieusement des attributs de données et des enregistrements les plus pertinents pour la prédiction. Ils peuvent être regroupes dans trois catégories :

• Attributs démographiques : contiennent les caractéristiques principales du client tels que le sexe, l'âge, la nationalité, le lieu de résidence, etc.

• Attributs du contrat : contiennent les attributs associes au contrat du client pour un service particulier tel que type de service, date de conclusion du contrat, prix du service etc.

• Attributs de comportement du client : décrivent l'activité du client.

Les données soumises à notre analyse couvrent une période de quatre mois du 01/01/2019 au 01/05/2019. Le choix de cet intervalle est justifié par la date d'indexation des numéros et la date d'exploitation des données par l'analyste.

La plupart des algorithmes d'exploration des données nécessitent une table qui englobe toutes les informations clients pour la modélisation, et pour cela, les données clients des deux tables, qui sont issues des sondes otarie et astellia, sont intégrées pour fournir une vue unique du client.

Cette table globale fusionne les informations liées à la clientèle grâce à une variable identifiant client : dans notre cas notre variable est le 'MSISDN', tout en dégageant des nouvelles variables appelées des kpi.

Pour répondre à nos deux problématiques, nous avons passe par deux niveaux pour l'extraction des kpi's ; du plus général au plus spécifique.

✣ NIVEAU GENERAL

Dans cette approche nous avons suivis le comportement des clients de manière générale telles que le volume des données mobiles consommées, le nombre des appels entrants, le nombre des appels sortants ainsi que le nombre de fois qu'ils ont appelé le service de réclamation. Mais pour notre deuxième projet le flag était directement le montant de recharge de l'utilisateurs donc il ne fallait pas faire l'erreur de le sortir deux fois.

✣ NIVEAU SPECIFIQUE

Dans cette démarche, nous nous enfonçons encore plus sur le plan comportemental de l'abonne. En effet, nous avons établis des mesures en termes de distances. Distance entre les appels (entrants/sortants) exprimée en jour pendant les quatre mois, de même distance entre les recharges effectuées ainsi que les recharge par rapport la date d'expédition pendant les quatre mois.

MSISDN	0
FLAG	0
TENURE	0
LAST_DISTANCE_ACHAT	0
LAST_DISTANCE_ENTRANT	0
LAST_EXPIRY_RECHARGE_DIST	0
LAST_DISTANCE_RECHARGE	0
LAST_DISTANCE_SORTANT	0
LAST_DISTANCE_SOS	0
LIFE_TIME	0
MIN_BTWACHAT	0
AVG_BTWACHAT	0
MAX_BTWACHAT	0
MIN_BTW_RECHARGE	0
AVG_BTW_RECHARGE	0
MAX_BTW_RECHARGE	0
MIN_BTWSOS	0
AVG_BTWSOS	0
MAX_BTWSOS	0
MONTANT_SOS	0
NB_CALL_CENTER	0
NB_APPEL_ENTRANT	0
NB_MINUTES	0
NB_ACHAT_OPTION	0
NB_RECHARGE	0
NB_APPEL_SORTANT	0
NB_SOS	0
MIN_BTW_APPEL_ENTRANT	0
AVG_BTW_APPEL_ENTRANT	0
MAX_BTW_APPEL_ENTRANT	0
MIN_BTW_APPEL_SORTANT	0
AVG_BTW_APPEL_SORTANT	0
MAX_BTW_APPEL_SORTANT	0
OFFRE	0

FIGURE : KPIS DU PROJET DE PREDICTION DE CHURN

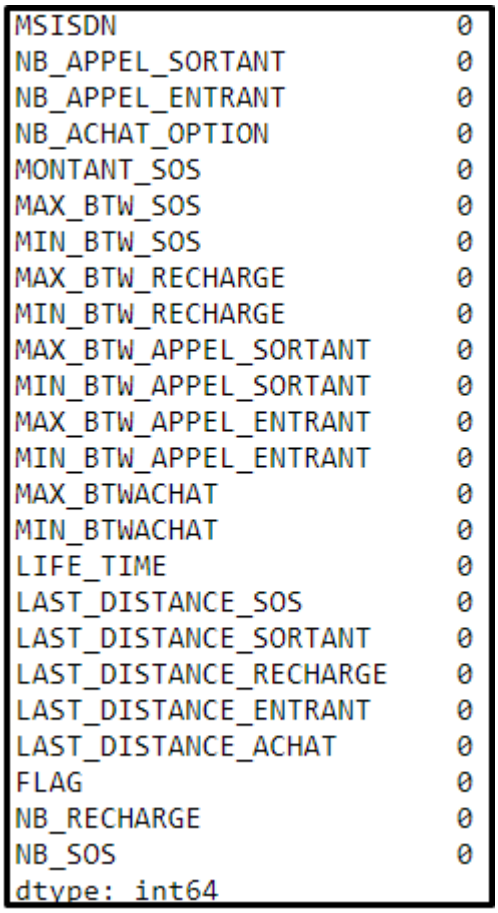


FIGURE : KPIS DU PROJET DE PREDICTION DU MONTANT DES RECHARGES

c) NETTOYAGE DES DONNEES

Lors de la préparation des données, nous avons remarqué la présence de bruit, de valeurs inconnues ou vides, de valeurs aberrantes et de valeurs non valides qui peuvent affecter négativement sur les performances des algorithmes d'apprentissage automatique en utilisant les données brutes.

Le nettoyage des données a pour but de réduire le nombre de valeurs incohérentes, de supprimer le bruit et les entrées et attributs incomplets. Notre ensemble de données étant suffisamment volumineux, nous avons supprimé tous les nuplets potentiellement problématiques. Mais pour quelques kpis on a décidé de remplacer la valeur nulle par la moyenne de la colonne concernée pour quand même garder le maximum de data.

Quand on vérifie aussi les valeurs des différentes kpis, il vaut mieux avoir des valeurs pas très distante et grande les unes des autres. Pour éviter cela, on a appliqué une normalisation sur les données. Le but de la normalisation est de changer les valeurs des colonnes numériques du jeu de données en une échelle commune, sans distorsion des

différences entre les plages de valeurs. Pour l'apprentissage automatique, chaque jeu de données ne nécessite pas de normalisation. Il n'est requis que lorsque les fonctionnalités ont des plages différentes.

	NB_ACHAT_OPTION	NB_APPEL_ENTRANT	NB_APPEL_SORTANT	NB_RECHARGE	NB_SOS
count	34887.00	34887.00	34887.00	34887.00	34887.00
mean	36.18	1636.56	1205.78	8.05	24.52
std	38.29	1450.54	1239.89	3.52	32.53
min	1.00	26.00	1.00	1.00	1.00
25%	10.00	772.00	437.00	6.00	4.00
50%	30.00	1275.00	870.00	7.00	12.00
75%	42.00	2044.00	1560.00	10.00	32.00
max	695.00	41639.00	25000.00	37.00	482.00

On peut voir ici qu'une normalisation est nécessaire étant donné que les valeurs « max » sont très différentes et possèdent un grand écart. Bien évidemment cela est valable pour les 2 projets. Toutes les étapes qui se font sur les données sont aussi valables pour les 2 projets.

Il existe une autre étape fondamentale pour le nettoyage des données qui est de vérifier les corrélations des kpis pour éviter tout risque d'overfitting ou l'underfitting.

- L'overfit est un surapprentissage, il survient lorsqu'un modèle apprend les détails et le bruit dans les données d'apprentissage dans la mesure où ils ont un impact négatif sur les performances du modèle avec de nouvelles données. Cela signifie que les variations de bruit ou aléatoires dans les données d'apprentissage sont captées et apprises en tant que concepts par le modèle. Le problème est que ces concepts ne s'appliquent pas aux nouvelles données et ont un impact négatif sur la capacite des modèles à généraliser.

- L'underfit est le sous-ajustement, il fait référence à un modèle qui ne peut ni modéliser les données d'apprentissage ni se généraliser a de nouvelles données. Un modèle d'apprentissage machine sous-équipe n'est pas un modèle approprie et sera évident car il aura de mauvaises performances sur les données d'entrainement. La solution consiste à essayer d'autres algorithmes d'apprentissage machine.

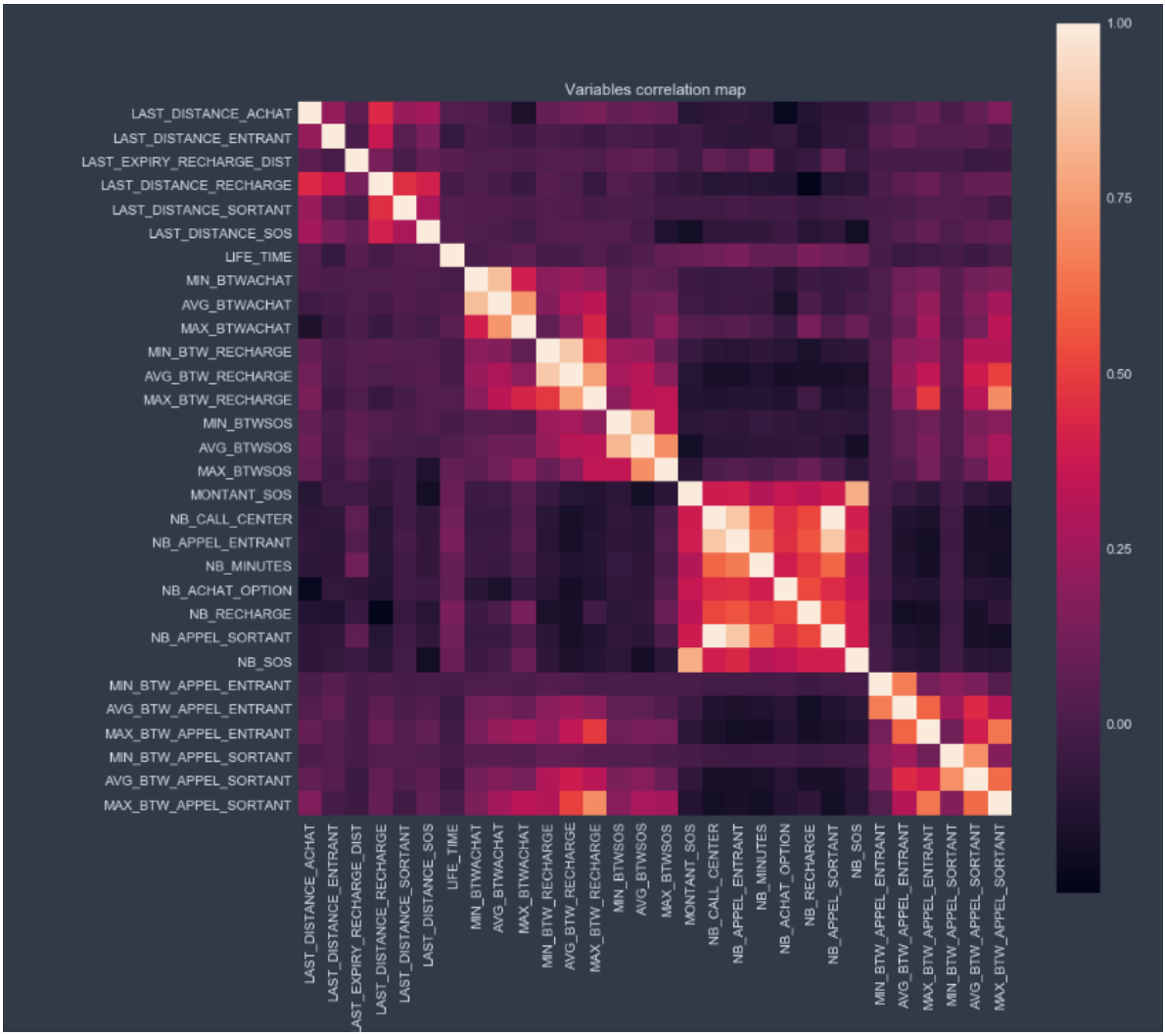


FIGURE : TABLE DE CORRELATION FINALE POUR LE PROJET DE CHURN

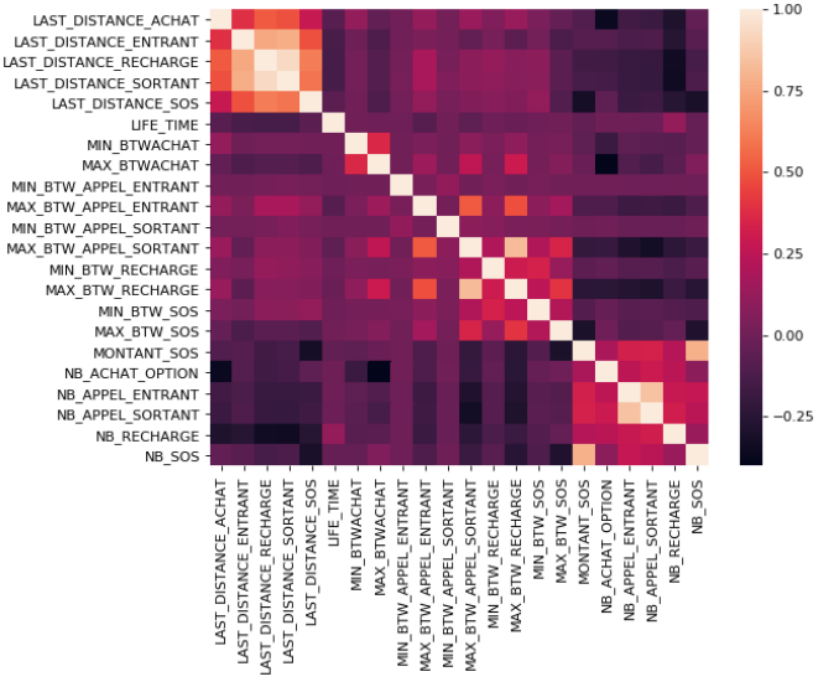
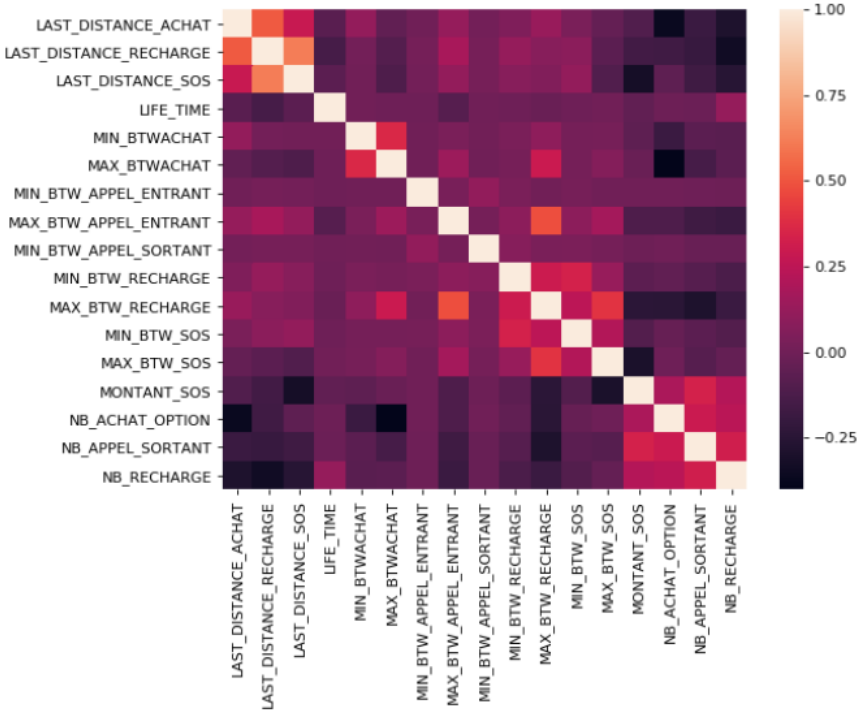


FIGURE : TABLE DE CORRELATION PROJET RECHARGE AVANT NETTOYAGE

Si on fait bien attention à la table de corrélation du projet de churn on pourra voir une forte corrélation entre nb_call_center et nb_appel_sortant. Ce qui est normal car quand on appelle le call center on effectue un appel sortant et on a juges avec mon collègue qu'on pouvait garder les 2 kpis car elles sont importantes toutes les deux.



Mais dans le deuxième projet, on possède une forte corrélation entre le nombre d'appel sortant et d'appel entrant. Ce qui est normal aussi mais dans ce cas-là, on sait que les appels sortants consomment des crédits et on peut dire qu'ils sont plus importants que les appels entrants. J'ai alors supprimé la kpi.

FIGURE : TABLE DE CORRELATION PROJET RECHARGE APRES NETTOYAGE

Durant les deux projets du stage, j'ai pu découvrir différentes façons de faire. Généralement les kpis avec une forte corrélation on en choisit une (la moins importante et qui a le moins d'impact entre les deux) pour la supprimer et enlever la corrélation mais parfois on doit les garder. Chaque ingénieur possède une façon de faire différente, il n'y a donc pas de règle générale a ce genre de problème mais plusieurs méthodes.

L'extraction, le calcul et le nettoyage des kpis du projet de churn a été beaucoup plus long et difficile que le projet des recharges bien que toutes les étapes étaient similaires. Cela est dû au fait qu'il y avait plus de variables à étudier et des kpis à avoir. Cette étape est généralement la plus chronophage de toutes les étapes du processus puisqu'elle représente environ 80% du travail d'un data scientist afin de préparer les données pour l'analyse et la visualisation.

III. ANALYSE DES DONNEES ET MODELISATION

Dans cette étape, nous créons et validons un modèle d'apprentissage automatique utilise dans le traitement de données. Le plus souvent, le modèle d'apprentissage automatique final est déployé dans le contexte d'une application pour fournir des informations utiles (telles que la classification ou la prédiction).

D'abord, nous ajustons un modèle en attribuant un ensemble de données d'apprentissage afin de l'entrainer et de faire des prédictions fiables sur des données nouvelles ou non vues.

Ensuite, nous validons le modèle forme au moyen de techniques de validation croisée qui permettent une bonne généralisation et évitent le surentrainement. L'idée est de diviser le jeu de données en deux sous-ensembles : un sous-ensemble est utilisé pour l'entrainement, tandis que l'autre pour un test et les performances du modèle final sont évaluées. La validation croisée vise principalement à obtenir une estimation stable et fiable des performances du modèle.

a) PROJET CHURN

Pour le projet de churn, on a utilisé la librairie sklearn pour séparer les données de test et de training.

Lors de l'exploration de nos données, nous avons constaté un grand déséquilibre. Ce problème est très fréquent dans la science des données et plus particulièrement dans les problèmes de **classification** et signifie que les classes ou les catégories que nous voulons prédire ne sont pas représentées de manière égale dans l'ensemble de l'entrainement. Cela peut conduire à prédire les classes dominantes plus que les autres, on a donc fait attention de les avoir similaire ou très proches.

L'un des moyens que nous n'avons pas évoqués dans le problème de classification lie à l'apprentissage automatique et qui pourrait résoudre le problème de données non équilibrées consiste à réduire le nombre d'échantillons étiquetés aux populations les moins dominantes. Ce déséquilibre peut être réduit dans une large mesure en sous-échantillonnant la classe majoritaire, étiquette 0, et en le rapprochant de celui de l'étiquette 1. Cette technique en anglais s'appelle Under-sampling.

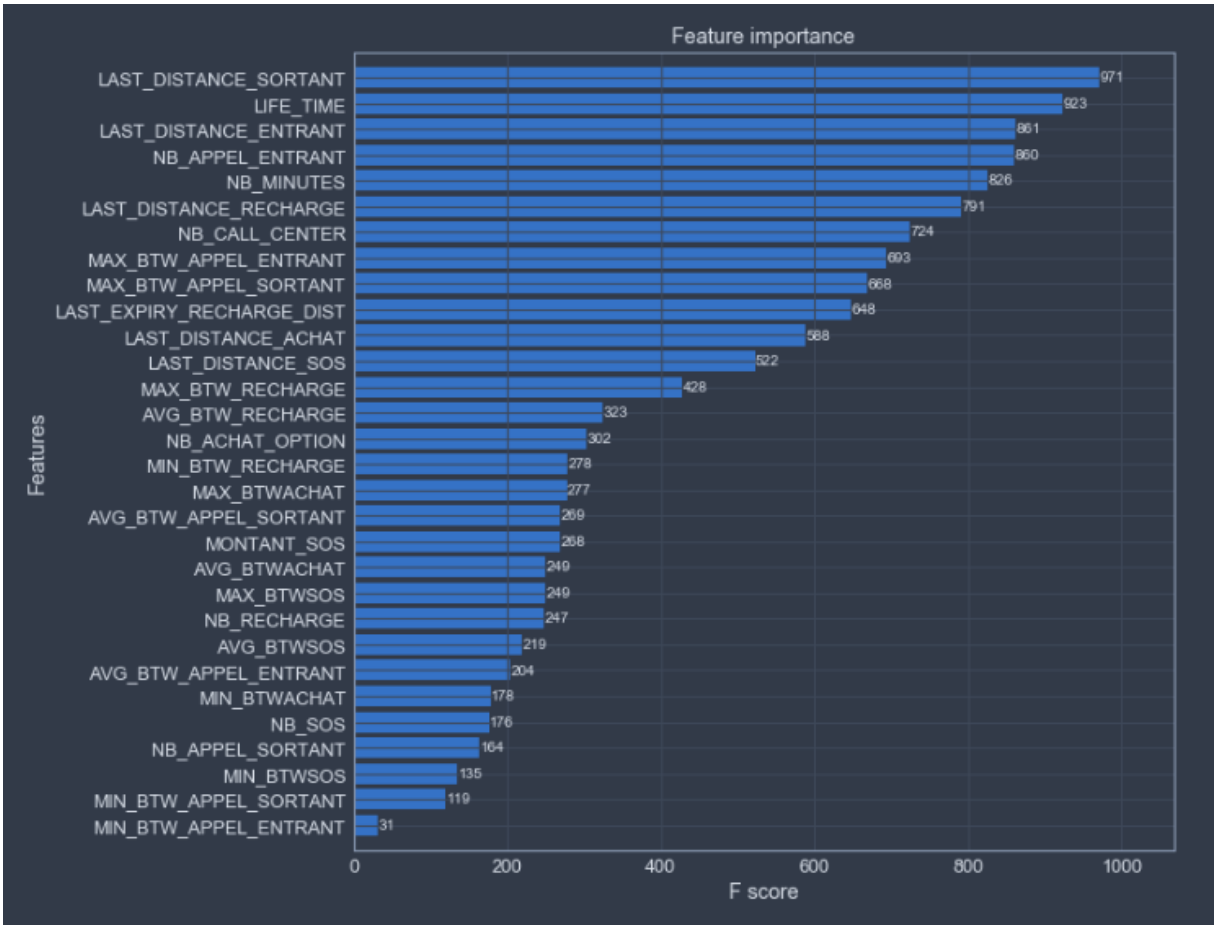
Nous divisons notre ensemble de données en un ensemble d'apprentissage (70% des données étiquetées disponibles) et un ensemble de test (30% des données étiquetées disponibles) mais un ensemble de données composes de nombre égale des deux classes. On utilise plusieurs algorithmes d'apprentissage et on les compare pour choisir le meilleur parmi eux :

	Model	accuracy
0	random forest	0.869249
5	gradient boosting	0.869124
9	Lightgbm	0.868998
8	Catboost	0.865864
7	XGB	0.864987
6	SVM	0.830889
4	KNN	0.807572
1	logistic regression	0.786762
3	naive bayes	0.690611
2	adaboost	0.626551

TABLEAU : RESULTAT DES DIFFERENTS ALGORITHMES D'APPRENTISSAGE

L'accuracy est un score qu'on obtient après la phase test. Quand notre programme s'entraine et ensuite se test sur les 30% restant on obtient alors un rapport de précision.

On vérifie ensuite l'impact de chaque kpi pour mieux comprendre notre modèle et avoir de possibles améliorations pour le futur. On a un « classement » de leur impact sur le modèle.



GRAPHE : FEATURE IMPORTANCE POUR MODEL DU PROJET CHURN

b) PROJET RECHARGE

Pour le projet de recharge, on va réaliser les mêmes étapes. C'est très similaire avec le churn a l'exception que cette fois ci ça sera un problème de **régression** et non pas de classification. La modélisation prédictive par régression consiste à approximer une fonction de mappage (f) de variables d'entrée (x) a une variable de sortie continue (y).

Une variable de sortie continue est une valeur réelle, telle qu'un entier ou une valeur a virgule flottante. Ce sont souvent des quantités, telles que des quantités et des tailles.

On a entrainé notre modèle pour qu'il nous prédit les montant de recharges des utilisateurs. Il y aura ici aussi la phase d'entrainement et la phase de test avec des données partager pour les 2 (70% de train et 30% de test).

Dans le cas d'une régression, on ne peut vérifier réellement avec l'accuracy car le modèle va prédire un montant qui est un float (un nombre a virgule).

Il existe de nombreuses façons d'estimer les compétences d'un modèle prédictif de régression, mais la plus courante consiste peut-être à calculer l'erreur quadratique moyenne, abrégée en acronyme RMSE.

RMSE est une règle de notation quadratique qui mesure également la magnitude moyenne de l'erreur. C'est la racine carrée de la moyenne des différences au carre entre la prévision et l'observation réelle.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Un avantage de RMSE est que les unités du score d'erreur sont dans les mêmes unités que la valeur prédite.

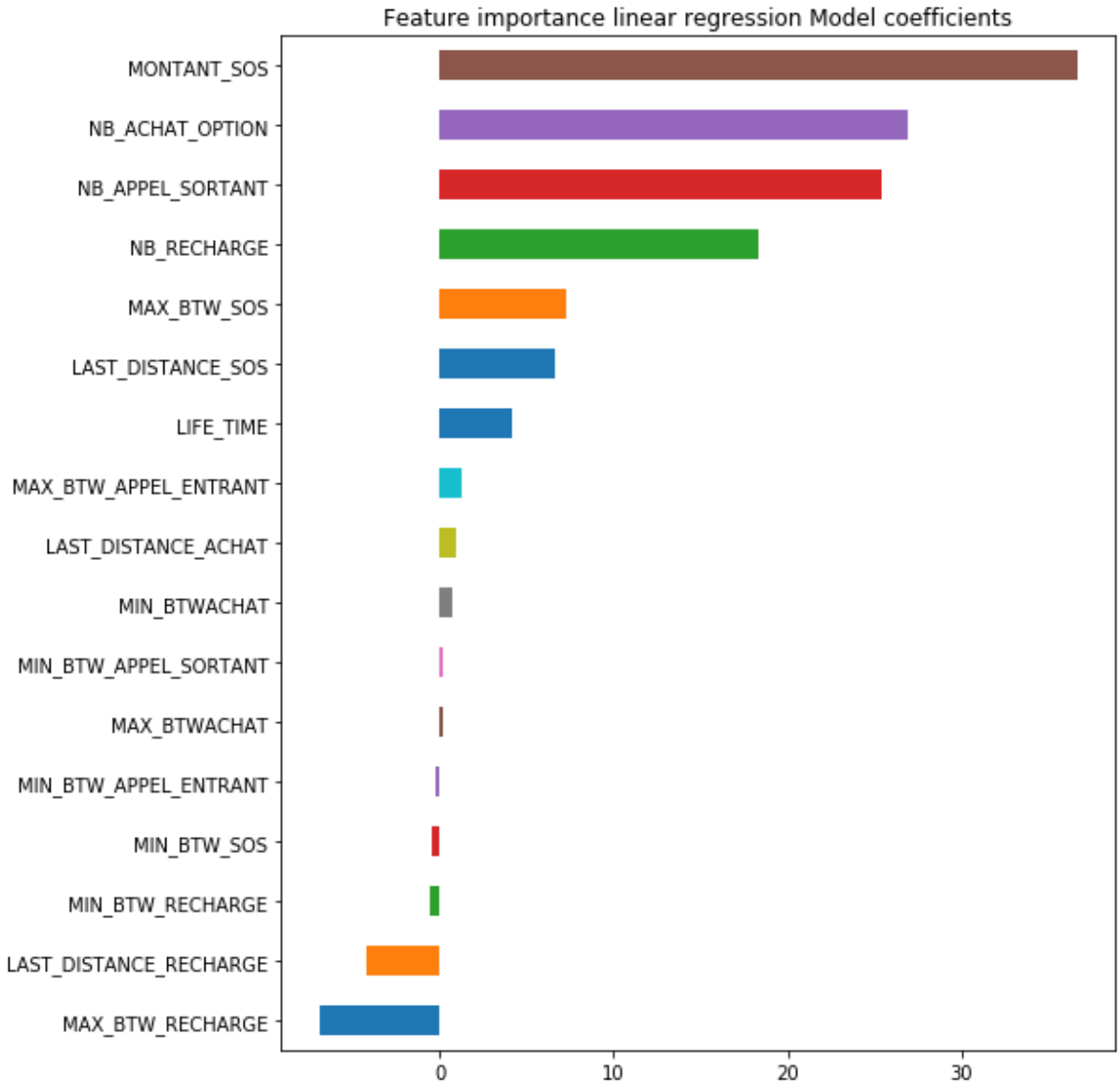
C'est ce qu'on a donc fait avec un modèle d'apprentissage qui est la régression linaire. On a obtenu un RMSE = 49 que mes supérieurs ont jugé suffisant pour valider le modèle.

Il faut savoir que les valeurs de RMSE sont difficilement interprétables car elles varient en fonction des problématiques.

Pour mieux comprendre cela on va prendre un exemple. Si on a un RMSE de 10, il est considéré relativement faible si la moyenne des observations est de 500. Pourtant, un modèle a une variance forte s'il conduit à un RMSE de 10 alors que la moyenne des observations est de 15.

En effet, dans le premier cas, la variance du modèle correspond à seulement 5% de la moyenne des observations alors que dans le second cas, la variance atteint plus de 65% de la moyenne des observations.

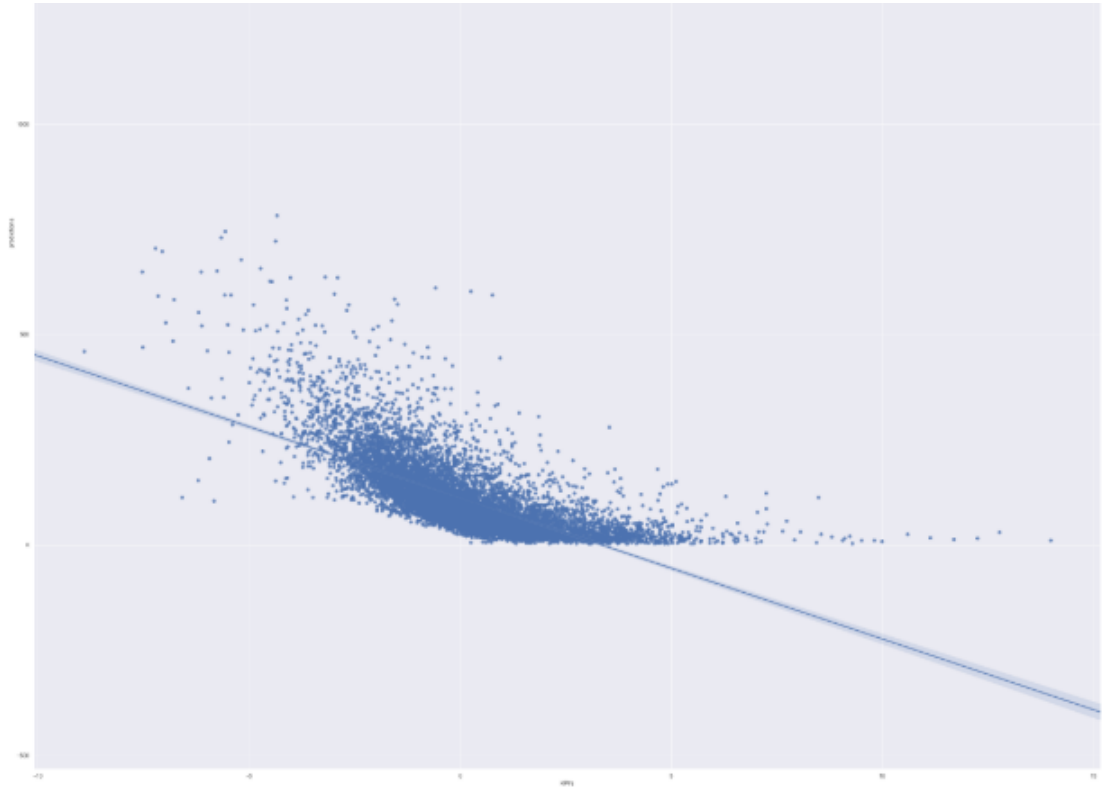
Après cela on étudie l'importance de chaque kpi ici aussi et on peut même tracer la droite de régression linaire en plein dans le nuage de point pour illustrer le modèle, tout cela grâce à des librairies tel que matplotlib et sns.



GRAPHE: FEATURE IMPORTANCE LINEAR REGRESSION MODEL

MAX_BTW_RECHARGE	-6.90
LAST_DISTANCE_RECHARGE	-4.22
MIN_BTW_RECHARGE	-0.53
MIN_BTW_SOS	-0.45
MIN_BTW_APPEL_ENTRANT	-0.19
MAX_BTWACHAT	0.20
MIN_BTW_APPEL_SORTANT	0.25
MIN_BTWACHAT	0.77
LAST_DISTANCE_ACHAT	1.00
MAX_BTW_APPEL_ENTRANT	1.27
LIFE_TIME	4.20
LAST_DISTANCE_SOS	6.61
MAX_BTW_SOS	7.28
NB_RECHARGE	18.35
NB_APPEL_SORTANT	25.45
NB_ACHAT_OPTION	26.87
MONTANT_SOS	36.60

FIGURE : LE COEFFICIENT DE L'IMPACT DE CHAQUE KPI



GRAPHE : NUAGE DE POINT DES DONNEES AVEC LA DROITE DU MODELE DE REGRESSION LINAIRE

Pour conclure cette partie, nous avons explore les objectifs atteints par notre projet de fin d'étude. Nous avons d'abord étudié les technologies choisies pour mettre en œuvre notre projet. Ensuite, élucider les différentes étapes aboutissant à la prédiction en passant en revue la préparation et l'analyse des données.

BILAN PERSONNEL

Pour commencer, je dirais que ce stage a été à la hauteur de mes attentes et m'a permis de compléter ma formation. Il m'a aussi permis à connaitre un nouveau domaine que je ne connaissais pas vraiment qui est le machine Learning. J'ai pu découvrir un secteur de travail qui est la télécommunication, qui est vraiment très intéressant et m'a agréablement surpris par sa complexité mais aussi de sa diversité et sa face cachée qu'on ne connait pas forcement en tant que simple utilisateur.

J'ai pu murir sur le plan technique mais aussi personnel. C'était un stage assez différent de ce qu'on faisait en cours et qui correspondait énormément à mon projet professionnel qui est de m'orienter vers la data science. A mon arrive à la boite, j'ai pu découvrir tous les domaines (front, back, big data). En tant qu'étudiant dans un master bi, j'ai découvert les deux facettes de ce domaine, mais dès que j'ai commencé à chercher le stage, je savais que je voulais travailler dans l'intelligence artificielle.

Dans un point de vue technique, j'ai pu apprendre rapidement l'utilisation du machine Learning avec le langage python et l'utilisation de plusieurs librairies différentes. Chaque librairie m'aidait à résoudre une nouvelle problématique a fur et à mesure que j'avançais dans le projet. A partir de la moitié du 2eme mois j'étais apte à travailler sur un projet de machine Learning seul. Mon collègue m'a beaucoup aidé et il a pu me corriger les erreurs fréquentes et me conseiller de comment les éviter lors d'un projet. Cela m'a été énormément bénéfique lors de la réalisation du projet de prédiction de recharge que j'ai pu faire seul.

Durant certaines journées, d'autres collègues me demandaient mon aide sur un travail différent du mien comme par exemple : modifier une table dans la DWH ou encore éditer un Dashboard à l'aide de qlisksense. Ces petites expériences m'ont permis de découvrir des logiciels différents qui concerne mon domaine et mes études, notamment comme Talend et qlisksense. En tant que futur ingénieur, j'ai aimé être dans cette position une personne polyvalente, je me suis parfaitement épanouie, je pense avoir fait un bon choix de stage.

Ce stage m'a permis de découvrir des compétences que je ne pensais pas avoir, comme une autonomie efficace. Le travail en groupe m'a été énormément bénéfique aussi, l'échange nous apprend beaucoup à toute l'équipe. Il y a de la confiance qui s'instaure une bonne ambiance au travail, on sait qu'on peut compter l'un a l'autre mais aussi un homme motive.

Des mes premières semaines, je me suis sentie intégrée et tout le monde m'a bien accueilli. Mon niveau relationnel s'est donc amélioré et j'ai pu prouver que je pouvais m'adapter à mon environnement comme j'ai pu le faire en arrivant en France et en partant en échange aux canadas. Je n'ai pas eu de problème de langue, je parlais parfaitement tunisien, et tous mes collègues parlait aussi le français et l'anglais. On a pu se rapprocher en organisant des activités hors du travail et cela m'a aidé à connaitre les personnes, et donc m'a indirectement aidé pour mon intégration et ma réussite au stage.

Je me suis personnellement senti dans un environnement professionnel, on avait des réunions chaque fin de semaine, j'ai appris à m'organiser et à prioriser les taches pour bien organiser mon temps, tout en restant flexible. Tout ce rythme de travail m'a un peu rappelé notre pa8 avec les réunions et les taches a réalisé. Cela m'a prouvé que tout ce qu'on faisait en cours pouvez réellement nous servir dans la vie professionnelle, bien plus qu'on le pensait.

Ma première mission a été la plus longue, bénéfique et enrichissante. On a passé beaucoup de temps dessus, et j'ai pu voir le modèle de classification en apprentissage supervise. Quant à ma 2e mission, elle m'a permis de me prouver que je pouvais travailler seul sur un projet de machine Learning de A à Z, et aussi que j'ai pu découvrir et travailler sur un modèle de régression en apprentissage supervise. Le machine Learning est vraiment intéressant et j'ai été heureux de travailler dessus, tout comme le domaine de télécommunication.

CONCLUSION

Pour conclure ce rapport, je dirais que ce stage a été très important pour mon avenir, c'était 20 semaine très enrichissantes et très productives pour moi. C'était ma première réelle expérience qui se rapproche le plus de mon domaine, j'ai eu la chance d'être très bien entourée le devoir eu un très bon environnement de travail. Cela a été vraiment bénéfique pour ma progression et mon apprentissage. C'est une expérience que je referai sans hésitation et dans le même domaine que j'ai réellement apprécié.

En commençant ce stage, je ne pensais pas atteindre ce niveau et de pouvoir apprendre autant de trucs en si peu de temps et pouvoir réaliser et mener à bien les deux projets. Mais tout s'est bien passe, mon tuteur de stage était fier de moi et j'étais fière de moi-même. Le fait d'avoir réussi les 2 projets et d'avoir appris le machine Learning, m'a été très bénéfique et je pense que cela sera très important pour mon avenir.

Aussi, ce stage m'a fait comprendre l'importance de la DSI dans un groupe. En tant qu'utilisateur on ne prend pas conscience de tout cela, mais une fois que vous travaillez dans ce service, c'est une toute autre vision que vous avez. Mon tuteur de stage a eu un énorme impact dans mon évolution, il m'a donné assez de liberté pour pouvoir tout découvrir. J'ai aussi réellement apprécié sa confiance en moi, je faisais attention à chacun de ses conseils car je savais que ça me sera utile.

Pour terminer, je pense que même pour mes futurs stages et cdi, je vais m'orienter vers la machine Learning et la data science. La méthodologie de travail et le rythme était vraiment excellent. Ce fut une bonne expérience et j'espère que mon stage de m2 sera tout aussi intéressant.

BIBLIOGRAPHIE

https://matplotlib.org/3.1.1/api/_as_gen/matplotlib.pyplot.plot.html

http://rasbt.github.io/mlxtend/user_guide/plotting/plot_linear_regression/

https://ledatascientist.com/creer-un-modele-de-regression-lineaire-avec-python/

https://docs.oracle.com/cd/B19306_01/server.102/b14200/functions137.htm

https://sql.sh/cours/merge

https://www.techonthenet.com/oracle/functions/coalesce.php

https://www.w3schools.com/sql/sql_update.asp

https://openclassrooms.com/fr/courses/4297166-realisez-des-calculs-distribues-sur-des-donnees-massives/4308656-familiarisez-vous-avec-hadoop

https://fr.coursera.org/learn/deep-neural-network

https://openclassrooms.com/fr/courses/4011851-initiez-vous-au-machine-learning/4020611-identifiez-les-differents-types-dapprentissage-automatiques

https://openclassrooms.com/fr/courses/4452741-decouvrez-les-librairies-python-pour-la-data-science

https://openclassrooms.com/fr/courses/4297211-evaluez-et-ameliorez-les-performances-dun-modele-de-machine-learning

https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/

LISTE DES SIGLES ET ACRONYMES

KPI: KEY PERFORMANCE INDICATOR.

INT: INSTANCE NATIONALE DES TELECOMMUNICATION.

BI: BUSINESS INTELLIGENCE.

ML: MACHINE LEARNING.

CDR: CALL DETAIL RECORD.

HUE: HADOOP USER EXPERIENCE.

MSISDN: NUMERO TELEPHONE DE L'ABONNE.

SQL: STRUCTURED QUERY LANGUAGE

CRM: CUSTOMER RELATIONSHIP MANAGEMENT.

MSC: MOBILE SWITCHING CENTER.

ERP: ENTERPRISE RESOURCE PLANNING.

BSCS: BUSINESS SUPPORT CONTROL SYSTEM.

DWH: DATA WAREHOUSE.

CSV: COMMA-SEPARATED VALUES.

ANNEXE

EXEMPLE CODE SQL LAST APPEL SORTANT :

```
select A.*, last_distance_appel_sortant
       from (select msisdn, sum(montant_recharge) montant_recharge from dmm_recharge
         where trunc(periode) > trunc(sysdate) - 120
         group by msisdn)A
left join (select s_p_number_address, last_distance_appel_sortant from(
    select s_p_number_address, abs(precedent_date - time_stamp) last_distance_appel_sortant from(
        select S_P_NUMBER_ADDRESS, time_stamp, precedent_date, precedent_S_P_NUMBER_ADDRESS from(
            SELECT S_P_NUMBER_ADDRESS, time_stamp, LAG(TIME_STAMP, 1) OVER (ORDER BY S_P_NUMBER_ADDRESS) precedent_date,
               LAG(S_P_NUMBER_ADDRESS, 1) OVER (ORDER BY S_P_NUMBER_ADDRESS, time_stamp) precedent_S_P_NUMBER_ADDRESS,
               row_number() over (partition by S_P_NUMBER_ADDRESS order by time_stamp desc) rank
            FROM(
                SELECT S_P_NUMBER_ADDRESS, TIME_STAMP,
                        row_number ()
                            OVER(PARTITION BY S_P_NUMBER_ADDRESS ORDER BY TIME_STAMP desc)
                FROM(
                    SELECT msisdn s_p_number_address, trunc(call_date) TIME_STAMP from REVENU_CONSO_PREPAID_MSISDN@link_bi
                          WHERE entry_date >= TRUNC (SYSDATE) - 120
                           AND trunc(call_date) >= TRUNC (SYSDATE) - 120
                            ))
                            )where S_P_NUMBER_ADDRESS = precedent_S_P_NUMBER_ADDRESS and rank = 1
                             order by S_P_NUMBER_ADDRESS
                            )
)where last_distance_appel_sortant is not null and last_distance_appel_sortant > 0)K
on A.msisdn=K.s_p_number_address
```

EXEMPLE CODE SQL NOMBRE APPEL ENTRANT :

```
select A.*, K.min_btw_appel_entrant, K.max_btw_appel_entrant
       from (select msisdn, sum(montant_recharge) montant_recharge from dmm_recharge
         where trunc(periode) > trunc(sysdate) - 120
         group by msisdn)A
left join (select s_p_number_address, min(diff) min_btw_appel_entrant, max(diff) max_btw_appel_entrant from(
    select s_p_number_address, abs(precedent_date - time_stamp) diff from(
        select S_P_NUMBER_ADDRESS, time_stamp, precedent_date, precedent_S_P_NUMBER_ADDRESS from(
            SELECT S_P_NUMBER_ADDRESS, time_stamp, LAG(TIME_STAMP, 1) OVER (ORDER BY S_P_NUMBER_ADDRESS) precedent_date,
               LAG(S_P_NUMBER_ADDRESS, 1) OVER (ORDER BY S_P_NUMBER_ADDRESS, time_stamp) precedent_S_P_NUMBER_ADDRESS
            FROM(
                SELECT S_P_NUMBER_ADDRESS, TIME_STAMP,
                        row_number ()
                            OVER(PARTITION BY S_P_NUMBER_ADDRESS ORDER BY TIME_STAMP desc)
                FROM(
                    SELECT s_p_number_address, trunc(start_time_timestamp) TIME_STAMP from KPI_DATA.DWH_INCOMING_VOICE
                          WHERE trunc(entry_date_timestamp) >= TRUNC (SYSDATE) - 120
                            AND trunc(start_time_timestamp) >= TRUNC (SYSDATE) - 120
                            ))
                            )where S_P_NUMBER_ADDRESS = precedent_S_P_NUMBER_ADDRESS
                            )
)where diff is not null and diff > 0
group by s_p_number_address)K
on A.msisdn=K.s_p_number_address
```

EXEMPLE CODE SQL NOMBRE ACHAT OPTION :

```
select A.*, nb_achat_option
       from (select msisdn, sum(montant_recharge) montant_recharge from dmm_recharge
         where trunc(periode) > trunc(sysdate) - 120
         group by msisdn)A
left join (select msisdn, sum(nombre) nb_achat_option
from DM_ACHAT_OPTION_MSISDN
WHERE entry_date >= TRUNC (SYSDATE) - 120 and
      date_achat >= TRUNC (SYSDATE) - 120
group by msisdn)K
on A.msisdn=K.msisdn
```