MEMOIRE DE STAGE DE FIN D'ETUDES

PREVISION DU TRAFIC VOIX POUR LE SERVICE VOICE INTERNATIONAL COCKPIT (VIC)

07/09/2020 – 06/09/2021

SY Abiboulaye

Étudiant en Master 2 - Data Science for Solutions Engineering

Alternant Data Scientist Junior chez Orange/OBS

Jury

Président du jury	Maitre d'apprentissage	Tutrice Pédagogique
BÉRARD Alexandre	Ivan Valiente	Myriam Bertrand

RESUME

Dans l'optique d'obtenir mon diplôme d'ingénieur à l'ESIEA, je devais réaliser une mission technique dans le cadre de mon alternance chez Orange/OBS (Orange Business Services) pendant la période du 06/02/2021 au 06/09/2021. Ma mission était sur la prévision du trafic voix pour le service Voice International Cockpit(VIC).
Orange Business Services offre à ses clients « Grandes Comtes Multinationaux » des services de téléphonie. Ces services font appel à des carriers locaux qui sont des fournisseurs d'accès de téléphonie

À l'ère du Big Data(données massives où les données sont de plus en plus volumineuses et arrivent sous différents formats avec une vélocité inouïe, il est nécessaire de pouvoir les analyser, explorer et éventuellement les modéliser pour pouvoir en tirer de la valeur. C'est dans ce cadre que j'ai été chargé d'explorer des modèles statistiques permettant d'estimer, de manière automatisée, le temps cumulé par mois des appels par pays et par carrier. Ce type de données sont des séries temporelles, c'est-à-dire des données collectées au cours du temps. Elles sont utilisées pour pouvoir décrire, analyser et faire des prédictions sur ses valeurs futures à partir de ses valeurs historiques
Pour bien réussir le projet, la première étape consister à identifier les besoins de l'expert métier et de prendre en considération plusieurs paramètres tels que : l'environnement technique du projet, le coût et le facteur humain. Cette première étape est l'étape primordiale qu'un data scientist doit cerner avant de commencer l'exploration
Plusieurs réunions ont été organisées pour présenter les travaux d'exploration aux experts métiers.
Cette mission m'a permis de développer des compétences techniques et humaines au-delà des compétences déjà acquises à l'école.

ABSTRACT

In order to obtain my engineering degree at ESIEA, I had to carry out a technical mission as part of my work-study at Orange/OBS (Orange Business Services) during the period from 06/02/2021 to 06/09/2021. My mission was on the prediction of voice traffic for the Voice International Cockpit (VIC) service. Orange Business Services offers its "Grandes Comtes Multinationaux" customers telephony services. These services use local carriers who are telephony service providers In the age of Big Data (big data where data is increasingly large and arrives in different formats with unprecedented velocity, it is necessary to be able to analyze, explore and possibly model it to be able to derive value from it. It is in this context that I was asked to explore statistical models to estimate, in an automated way, the cumulative time per month of calls by country and by carrier. This type of data is time series, i.e. data collected over time. They are used to be able to describe, analyze and make predictions about its future values from its historical values to succeed in the project, the first step is to identify the needs of the business expert and take into consideration several parameters such as: the technical environment of the project, the cost and the human factor. This first step is the essential step that a data scientist must identify before starting the exploration. Several meetings were organized to present the exploration work to the business experts. This mission allowed me to develop technical and human skills beyond the skills already acquired at school.

SOMMAIRE

Table des Illustrations

REMERCIEMENTS

Je tiens à remercier tous ceux qui, de près ou de loin, ont contribué à la réalisation de ce projet :
Toute l'équipe d'Orange/OBS (Orange Business Services) pour leur accueil chaleureux et leur ouverture d'esprit qui m'ont permis de réaliser mon alternance dans les meilleures conditions possibles.
L'ESIEA pour le fait de m'avoir formé aux différentes compétences technique et humaine qu'un ingénieur doit avoir.
Mon tuteur d'apprentissage Ivan Valiente pour son accueil et sa disponibilité tout au long du projet. Je le remercie également de m'avoir formé au métier de data scientist dans un contexte business. J'en profite aussi pour remercier toute l'équipe notamment mon manager Valérie Marchand, Catherine Audusseau et Sebastien Praquin qui ont proposé le sujet et m'ont permis de pouvoir travailler sur une notion qui m'était inconnu et tous les autres membres de l'équipe qui m'ont aidé lors de mon projet.
Ma tutrice pédagogique Myriam Bertrand qui m'a fourni les documents nécessaires pour pouvoir comprendre les séries temporelles, une notion technique que j'avais besoin d'assimiler pour pouvoir traiter mon sujet.
Mes parents pour leur soutien sans failles ainsi que toute ma famille et amis.

INTODUCTION ET CONTEXTE DU STAGE

1 .1 Présentation de l'entreprise

Figure 1 : Groupe Orange en chiffres

Le Groupe Orange est une entreprise multinationale présente dans 24 pays et a plus de 147 000 employés. Le PDG actuel du groupe est Stéphane Richard. Il est considéré comme 8^{e} marque mondiale de télécommunication et est un acteur de confiance qui donne à chacun et à chacune les clés d'un monde numérique responsable. Le chiffre d'affaires de l'entreprise est estimé à 42.2 milliards.

Le groupe Orange donne à ses collaborateurs les moyens d'agir pour :

- Une économie responsable
- L'égalité numérique
- Une société de confiance
- L'environnement

1.2 Plan stratégique du groupe

La data et l'IA (Intelligence Artificielle) est un des axes stratégiques du groupe.

Figure 2: Programme Orange Learning

Cela explique le fait que le groupe

1.4 Présentation de la filiale OBS et l'entité Data & Billing

Figure 3Entité Data&Billing

Au sein d'Orange, existe plusieurs filiales mais la filiale **OBS**(Orange Business Servivces) est la filiale sur laquelle j'ai effectué mon alternance.Sur la filiale on trouve dans l'ordre :

- L'entité **CMI** ou Customer Marketing Innovation
- L'entité **OBS-IT** où se trouve tout ce qui est lié aux activités IT(Information Technology)

- L’entité **DB** (Data & Billing) est une ESN(entreprise de services numériques) , spécialisé dans l’IA(Intelligence Artificielle) , au sein duquel on retrouve l’entité **Analytics & Data Intelligence**. Son but est d’accompagner les entreprises dans la digitalisation de leurs processus métiers et en les aidant à mieux valoriser leurs données.

PRESENTATION DU PROJET

Contexte et Objectifs du projet

Orange Business Services offre à ses clients « Grandes Comtes Multinationaux » des services de téléphonie. Ces services font appel à des carriers locaux. Les carriers sont des entités fournissant des services de télécommunications (fixes, mobiles, voix IP) en tant qu'activité principale à l'ensemble ou à un sous-ensemble de consommateurs, entreprises, gouvernements et autres fournisseurs de services de télécommunications.

Aujourd'hui les équipes d'OBS ne disposent pas d'outils automatisés qui permettraient d'élaborer le budget prévisionnel, c'est-à-dire les moyens d'anticiper les charges par carrier dans les différents pays.

L'objectif de ce travail est d'explorer des modèles statistiques permettant d'estimer, de manière automatisée, le temps cumulé par mois des appels par pays et par carrier. Cette estimation de durée cumulée des appels sera utilisée à l'élaboration du budget prévisionnel

Les variables catégorielles que nous allons prendre en compte dans l'estimation de la durée cumulée des appels sont : ***le mois, le client, le type de service, le carrier (le fournisseur d'accès téléphonique local) et le pays.***

2. État de l'art

2.1. Recherche bibliographique sur les séries temporelles commentée

De nos jours, la quantité de données présentes dans plusieurs domaines tels que la finance, l'économie, la santé ainsi que dans le domaine de la télécommunication augmentent de manière exponentielle. La plupart d'entre elles sont des séries temporelles, c'est-à-dire des données collectées au cours du temps. Elles sont utilisées pour pouvoir décrire, analyser et faire des prédictions sur ses valeurs futures à partir de ses valeurs historiques. Concernant les méthodes de prédiction , d'importantes modèles à la fois linéaires et non linéaires sont utilisés de nos jours.

C'est pour cela que nous nous sommes intéressés aux références ci-dessous :

- (Éric Biernat) c'est un livre qui présente plusieurs méthodes de machine learning(apprentissage automatique). Il consacre trois chapitres sur les méthodes de séries temporelles et les méthodes linéaires utilsés actuellement

- (Sadigov)**:** c'est la formation proposée par Coursera sur les séries temporelles. On y retrouve les notions de **stationnarité,** la fonction d'autocorrélation pour déterminer l'ordre d'un processus MA(Moving Average), la fonction d'autocorrélation partiel pour déterminer l'ordre d'un processus AR(AutoRegressive), les modèles linéaires tels que le SARIMA(Seasonal AutoRegressive Integration Moving Average) et les méthodes de lissage exponentiel
- (Fatoumata Dama)

2.2. Analyse approfondie et critique des sources retenues

a- Définition d'une série temporelle

L'application $X : S \rightarrow \mathbb{R}$ est définie comme étant une variable aléatoire à valeurs réelles où S est l'espace d'expérimentation et $\mathbb{R}$ est l'espace d'arrivée.

Exemple :

Une urne contient trois boules de couleurs jaune(J), rouge(R) et verte(V). En faisant le tirage de deux boules successivement et en remettant la boule à chaque fois, les tirages possibles sont les suivants :

Tirage 1 / Tirage2	J	R	V
J	(J,J)	(R,J)	(V,J)
R	(J,R)	(R,R)	(V,R)
V	(J,V)	(R,V)	(V,V)

Dans notre cas , l'ensemble d'expérimentation ou l'univers de tous les possibles a pour cardinal 9. C'est l'ensemble des couples contenues dans le tableau ci-dessus.

Dans le cas d'un jeu dans lequel :

-Une boule J tirée fait gagner 2 euros

-Une boule R tirée fait gagner 1 euros

-Une boule verte perdre 2 euros

L'application X de S dans $\mathbb{R}$ qui, à tout tirage, associe le gain obtenu est appelée variable aléatoire :

Gain obtenu	J	R	V
J	4 euros	3 euros	0 euros
R	3 euros	2 euros	-1 euros
V	0 euros	-1 euros	-4 euros

Un **processus stochastique** est une collection ou famille de variables aléatoires X_1, X_2…. Dans le cas où ces variables aléatoires constituent une suite d'observations $(X_t)_{t\in S}$ et sont indicées par le temps, on parle de **séries temporelles** ou **séries chronologiques.** Une **série temporelle** est la réalisation d'un **processus stochastique**

L'ensemble S est appelé **espace de temps** dans le cas des séries temporelles et peut être soit:

- discret dans le cas où la date d'observation est équidistante comme c'est le cas du nombre de morts par jour dans l'épidémie du covid
- continu comme le cas des données météorologiques

b- Stationnarité des séries temporelles

L'étude de la stationnarité est une des étapes fondamentales avant de faire des prédictions sur les modèles linéaires tels que SARIMA

Un processus stochastique X est stationnaire si ses propriétés statistiques (la moyenne, la variance, autocorrélation) sont indépendantes du temps (Fatoumata Dama)

On distingue deux types de stationnarité d'après (Fatoumata Dama) :

- Stationnarité forte : Un processus stochastique X est de forte stationnarité si sa distribution satisfait la propriété suivante :
  $P(X_1, X_2, ..., X_T) = P(X_{1+\tau}, X_{2+\tau}, ..., X_{T+\tau}), \forall T, \tau \in N^*$
  Cela équivaut à dire que sa distribution est invariante à n'importe quel décalage horaire. Ce cas est rarement présent dans la pratique
- Stationnarité faible (weak stationnarity) : Un processus stochastique X est de faible stationnarité si :
  $E[X_t] = \mu$ (stationnarité de la moyenne)
  $Cov(X_t, X_{t+h}) = \gamma(h) \Longrightarrow Var(X_t) = \gamma(0) < \infty$ (stationnarité de la covariance)

Cela signifie que la moyenne ne dépend pas du temps i.e. qu'il n'y a pas de tendance. La fonction Cov traduit la fonction d'autocorrélation et le fait que la variance (Var) soit constante implique qu'il n'y a pas de changement dans la variation. Et d'après (Sadigov), la série ne doit pas avoir de saisonnalité pour qu'elle soit stationnaire.

Par la suite, à chaque fois que nous serons confrontés à une série qui n'est pas stationnaire , il faudra faire des transformations pour rendre la série stationnaire.

D'après (Fatoumata Dama)et (Éric Biernat) il existe deux méthodes pour tester la stationnarité de la série : la méthode graphique et la méthode statistique.

- Méthode graphique :

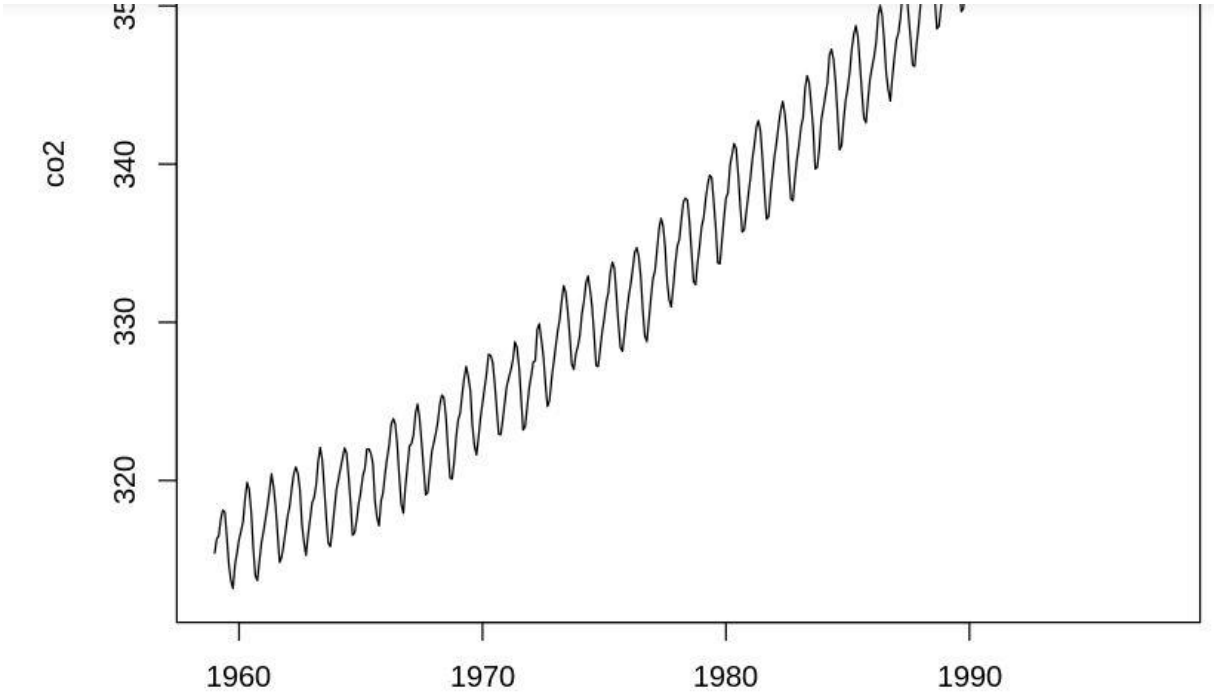

Figure 4: Augmentation co2 au fil des années

Dans le cas de la série qui est représentée sur la figure 4 et qui montre l'évolution de la quantité de co2 de 1960 à 1990, nous constatons qu'il y a une tendance croissante et une saisonnalité annuelle, ce qui montre que notre série n'est pas stationnaire

- Méthode statistique

On peut utiliser le test de Dickey-Fuller augmenté ou le test de racine unitaire

Le test de racine unitaire consiste à tester les hypothèses suivantes :

H0 : la série n'est pas stationnaire

Contre

H1 : la série est stationnaire

- Si la p-valeur est inférieure ou égale à alpha = 5%, le test est significatif au seuil alpha =5%. Nous rejetons H0 au seuil alpha=5% et nous décidons que H1 est vraie au seuil alpha =5% avec un risque d'erreur de première espèce alpha

- Si la p-valeur est supérieure à alpha =5%, le test n'est pas significatif au seuil alpha= 5%. Nous rejetons H1(l'hypothèse de stationnarité) et nous décidons que H0 est vraie au seuil alpha =5% avec un risque d'erreur de deuxième espèce alpha

c- Modèles linéaires : SARIMA

SARIMA(p,d,q,P,D,Q)$_s$ pour Seasonal AutoRegressive Modele Average est utilisé pour modéliser un processus ARIMA qui présente une saisonnalité.

D'après (Sadigov), il présente deux parties :

- Une partie qui ne présente pas de saisonnalité : p(ordre du processus AR), d(ordre de la différence),q(ordre du Moving Average)
- Une partie saisonnière : P(ordre saisonnier du processus AR), D(ordre saisonnier de la différentiation saisonnière), et Q(ordre saisonnier du processus MA)

Les étapes pour procéder à une modélisation sont les suivantes :

- Tracé et analyse de la série

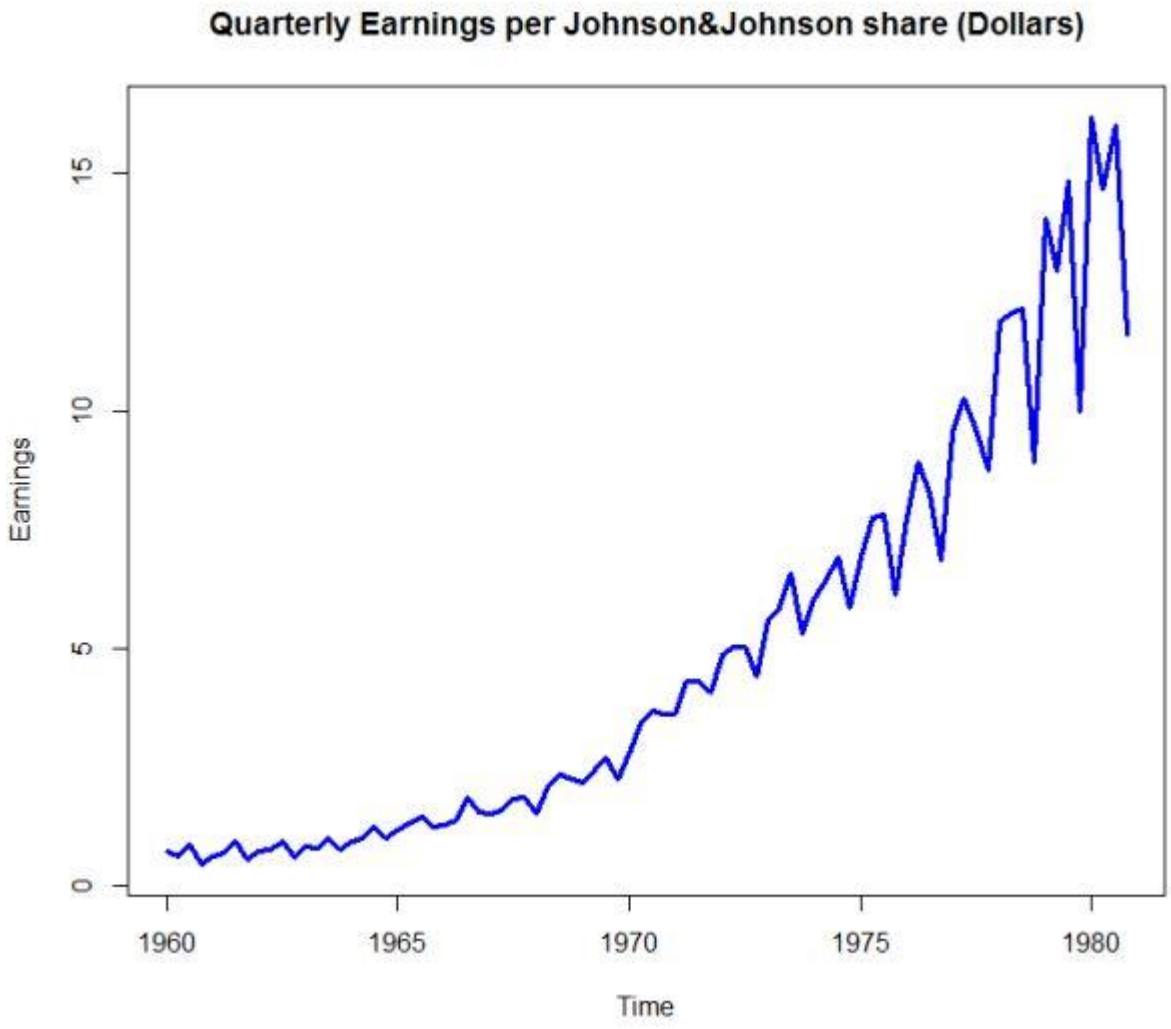

Figure 5 :Gain trimestriel Johnson&Johnson

La série ci-dessus représente le gain trimestriel obtenu par l'entreprise Johnson&Johnson de 1960 à 1980

- Transformation
  On observe que la série présente une tendance et que la variance n'est pas constante. On effectue quelques transformations pour la rendre stationnaire
- Obtenir la fonction d'autocorrélation pour déterminer q et Q
- Obtenir la fonction d'autocorrélation partielle pour déterminer p et P
- Entrainer différents modèles puis choisir celui qui a le plus faible AIC(Akaike Information Criterion)
- Principe de parcimonie : $p + d + q + P + D + Q \leq 6$

On observe les prédictions sur les trimestres suivants avec les intervalles de confiance à 80% puis à 95%. Nous pouvons conclure que notre modèle fait de bonnes productions

	Point for.	Lo 80	Hi 80	Lo 95	Hi 95
1981 Q1	2.910254	2.796250	3.024258	2.735900	3.084608
1981 Q2	2.817218	2.697507	2.936929	2.634135	3.000300
1981 Q3	2.920738	2.795580	3.045896	2.729325	3.112151
1981 Q4	2.574797	2.444419	2.705175	2.375401	2.774194
1982 Q1	3.041247	2.868176	3.214317	2.776559	3.305934
1982 Q2	2.946224	2.762623	3.129824	2.665431	3.227016
1982 Q3	3.044757	2.851198	3.238316	2.748735	3.340780
1982 Q4	2.706534	2.503505	2.909564	2.396028	3.017041

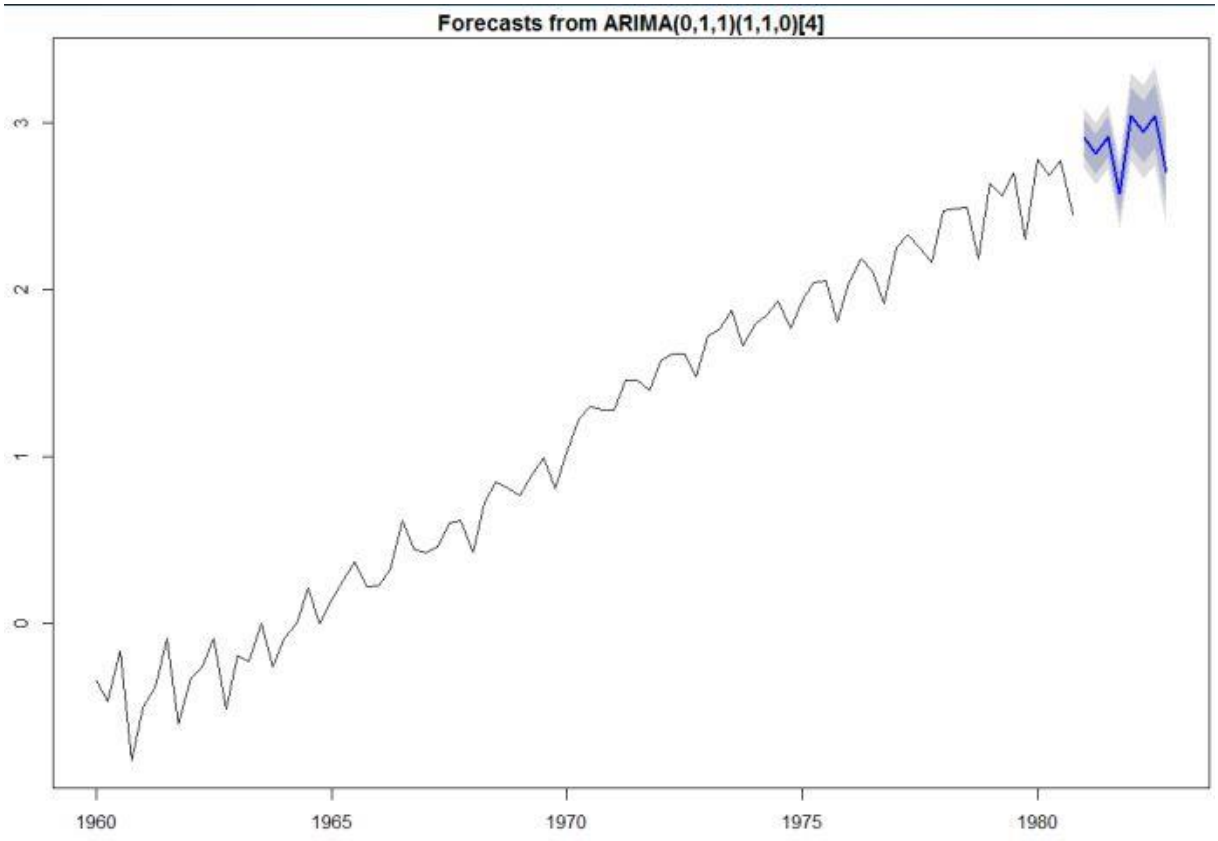

Figure 6: Prédiction pour les années 1981 et 1982

2.3. Synthèse

Nous pouvons conclure d'après les références bibliographiques que l'étude de la stationnarité de la série est très importante avant de procéder à la modélisation. Pour (Fatoumata Dama), le test de racine unitaire est nécessaire avant de passer à la modélisation tandis que pour (Sadigov) cette étape n'est pas nécessaire. La démarche de (Fatoumata Dama) (Fatoumata Dama) est beaucoup plus rigoureuse.

3. Dimensions techniques

Les technologies utilisées dans le cadre du projet sont les suivantes :

- Apache Hive qui est un data warehouse(entrepôt de données) intégré sur Hadoop permettant le requêtage des tables « Big Data »

- Python qui sont des langages de programmation les plus utilisés dans les études « data science »

- Modèles de séries temporelles : elles sont utilisées pour estimer l'évolution d'une variable au cours du temps à partir de ses valeurs historiques ; dans ce cas il s'agit d'anticiper le volume d'appels dans les mois suivants.

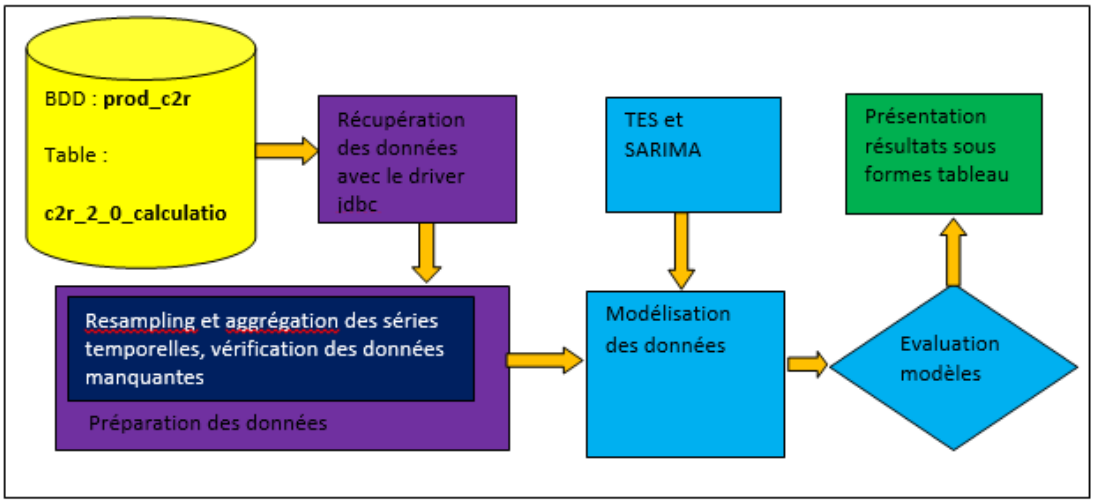

Figure 7 : Schéma détaillé du processus de traitement

3.1 Description de la source des données

La source de données est l'Enterprise Data Hub (EDH) dans laquelle se trouve la base de données (BDD) **prod_c2r** présentées dans le schéma de la figure 4. L'EDH permet la gestion des données massives (Big Data) en utilisant la plateforme Hadoop et permettant de faire
Les données peuvent être extraites manuellement soit via Hue soit via des scripts.

3.1.1 Récupération des données via Hue

Hue est une interface web contenant Hadoop et son écosystème(YARN, PIG, Hive etc) et permet d'interagir avec l'écosystème de manière interactive. Il est possible de faire des requêtes sur Hive avec le langage HiveQL(langage utilisé dans Hive pour faire des requêtes dans Hive) est proche du SQL (Structured Query Language) pour pouvoir récupérer les données et les télécharger en fichiers csv.

```
set hive.tez.container.size=8192;
set tez.queue.name=CostToRevenue;
SELECT distinct c2r_safran_id,c2r_safran_call_start_date, c2r_safran_cc_duration_min,c2r_inbound_cc_service_type, c2r_inbound_cost_eur,
c2r_inbound_carrier_name
FROM c2r_calculation WHERE c2r_safran_orig_country_code = "IND" AND c2r_safran_orig_access_type="MOBILE"
AND c2r_safran_call_start_date between "2021-03-31" AND "2021-05-01";
```

c2r_safran_call_start_date	c2r_safran_cc_duration_min	c2r_inbound_cc_service_type	c2r_inbound_cost_eur	c2r_inbound_carrier_name
2021-04-02 13:12:41.0	0.333	SAN INTERNATIONAL TOLL FREE SERVICE (ITFS)	0.02821	ORANGE IC (FORMERLY FT FREEPHONE) - FRANCE
2021-04-12 04:35:38.0	0.516	SAN INTERNATIONAL TOLL FREE SERVICE (ITFS)	0.06721	ORANGE IC (FORMERLY FT FREEPHONE) - FRANCE
2021-04-11 06:03:10.0	7.4	SAN INTERNATIONAL TOLL FREE SERVICE (ITFS)	1.554	BT (GSK) - IRELAND
2021-04-26 06:26:29.0	8.283	SAN INTERNATIONAL TOLL FREE SERVICE (ITFS)	NULL	BT - FRANCE (ACCOUNT 94635010)
2021-04-16 15:29:41.0	0.783	SAN INTERNATIONAL TOLL FREE SERVICE (ITFS)	0.16443	BT (GSK) - IRELAND

Figure 8: Extraction des données à travers Hue

Chaque requête sur Hue fournit au plus 1 million de lignes, Cette contrainte étant incompatible avec le volume de données nécessaires à l'estimation des modèles et à une future automatisation du processus de prédiction, nous a conduit à considérer une extraction semi-automatique via de scripts python.

3.1.2 Récupération des données à travers le driver jdbc

Pour faire face aux limites de récupération des données manuellement, une solution était d'automatiser la récupération des données à travers une API Jaydebe.

Le script d'extraction utilise le pilote jdbc et l'API disponible dans les modules python jpype et jaydebe.

Les colonnes de la table « c2r_2_0_calculation … » que nous avons utilisées dans cette exploration se trouvent dans le tableau suivant :

Variables	Interprétation	Data types
c2r_hookah_call_start_date :	Date de l'appel qui sera notre variable explicative pour prédire la durée	Timestamp
c2r_hookah_orig_access_type	Type d'accès: mobile ou fix	String
c2r_hookah_duration_min	La durée de l'appel, variable qu'on cherche à prédire	Float
c2r_inbound_cost_eur	Le coût de l'appel	Float
c2r_product_name	Le nom du produit	String

c2r_inbound_carrier_name	Le nom du carrier	String
c2r_hookah_orig_country_name	Le pays d'origine de l'appel	String

3.2 Analyse exploratoire des données

3.2 .1 Traitement des données

Les intervalles de temps entre deux valeurs successives de la colonne « c2r_hookah_call_start_date » n'étant pas constants, nous avons construit des séries temporelles par ré-échantillonnage avec une période journalière et en aggrégeant la somme des durées des appels de la journée comme le montre la figure 6

```
Entrée [11]: duree_jour_ind_fix.head(50)
```

Out[11]:

c2r_safran_call_start_date	c2r_safran_cc_duration_min
2020-12-30	699.956
2020-12-31	46969.929
2021-01-01	9944.340
2021-01-02	44730.493
2021-01-03	11835.411
2021-01-04	68604.512
2021-01-05	73891.402
2021-01-06	75786.513
2021-01-07	72680.758
2021-01-08	64033.689
2021-01-09	34099.642
2021-01-10	10419.258
2021-01-11	78739.802
2021-01-12	75402.480
2021-01-13	69617.689
2021-01-14	51624.850
2021-01-15	78732.673
2021-01-16	34481.355

Figure 9 Rééchantillonnage et agrégation de la durée

3.2.2 Base d'entrainement et base de test

Une fois que le rééchantillonnage et l'agrégation ont été faites, le dataframe est séparé en deux parties :

- Une partie pour l'entrainement du modèle : allant du 1er Janvier 2021 au 28 Février 2021
- Une partie pour tester le modèle sur le mois suivant qui lui est inconnu, dans notre cas le mois de Mars 2021

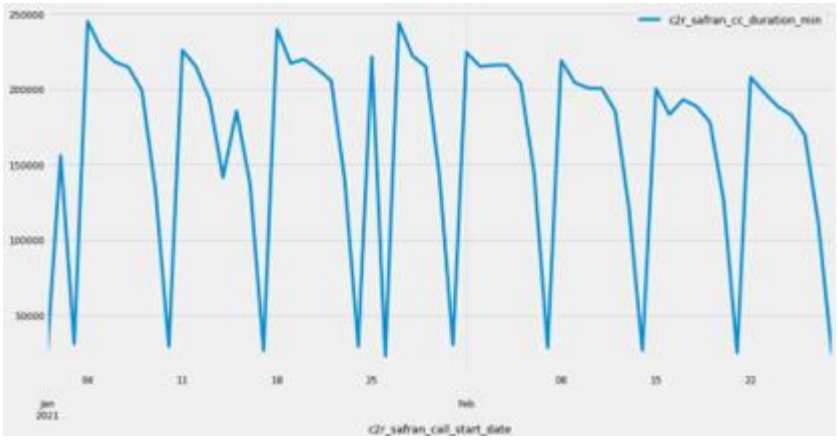

Figure 10 Base d'entraînement

Nous pouvons constater que nous avons une saisonnalité hebdomadaire sur la figure représentant la base d'entrainement.

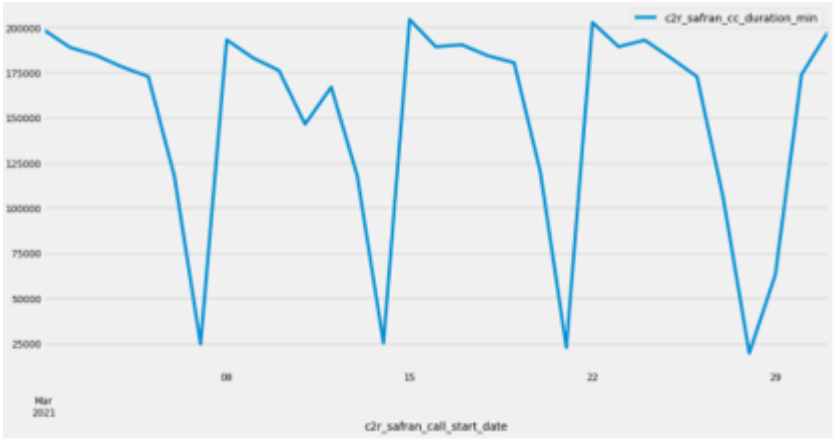

Figure 11 Base de test

3.2.3 Modélisation des données

Les modèles qui ont été testés sont les suivants :

- Modèle Seasonal Auto Regressive Integration Moving Average (SARIMA) de paramètre (p,d,q,P,D,Q,S)
- Modèle Single Exponential Smoothing (SES) de paramètre alpha : coefficient du level
- Modèle Triple Exponential Smoothing(TES) de paramètre alpha, beta et gamma où alpha, beta, gamma représentent respectivement les coefficients du level, du trend et de la

saisonnalité

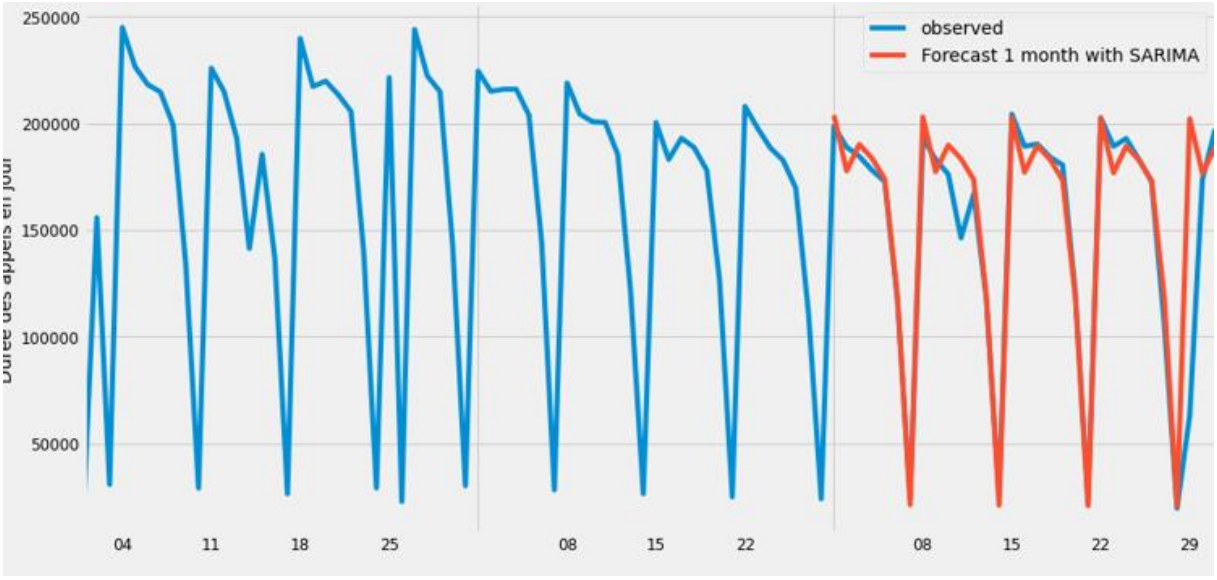

Figure 12 Modèle SARIMA pour la catégorie mobile du pays Inde

Fig : Modèle SARIMA pour la catégorie mobile du pays Inde

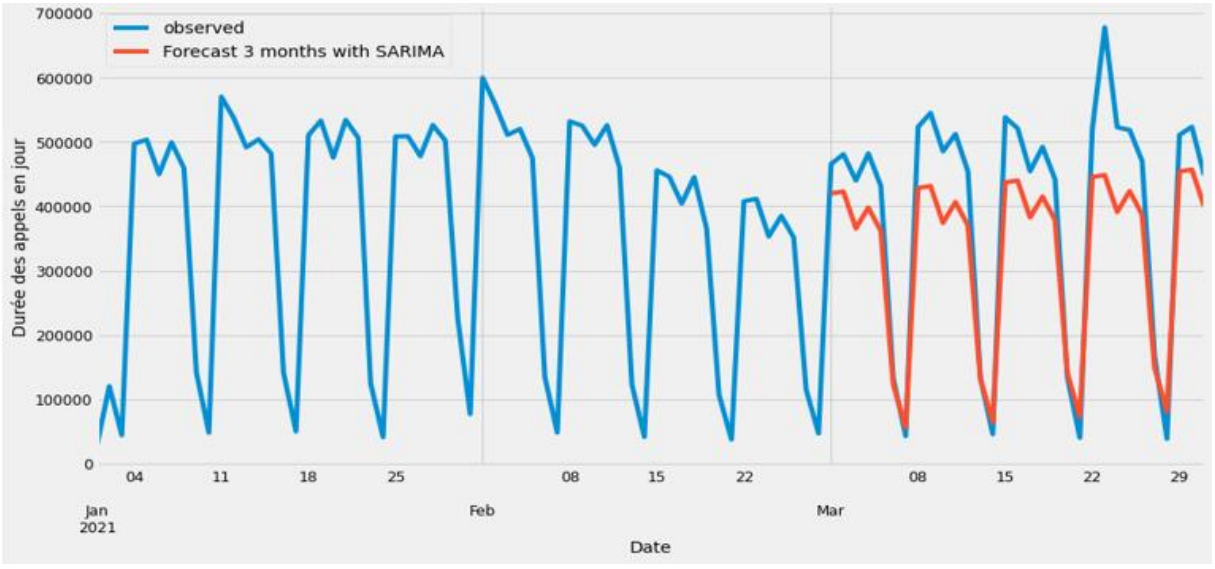

Figure 13 : modèle SARIMA pour la catégorie MOBILE du pays France

La partie en bleu représente le courbe réel et la partie en rouge représente la durée de la valeur prédite par le modèle SARIMA.

Pour le modèle SARIMA et TES, nous avons travaillé avec une saisonnalité hebdomadaire car nous avons limité l'historique de données à 122 jours pour des raisons du volume de données à traiter.

L'objectif était de pouvoir faire une prédiction de la durée cumulée des appels sur un mois avec un avec un seuil d'acceptabilité de 15% d'erreur relative.

Les conditions d'évaluation des modèles sont les suivantes:

a) 3 mois d'apprentissage (estimation des paramètres des modèles), et prédiction du quatrième mois disponible dans le dataset.

b) Evaluation de l'erreur relative entre la durée cumulée du mois prédite et la durée cumulée réelle du quatrième mois.

c) Si l'erreur relative du modèle est inférieure au seuil d'acceptabilité, le modèle est considéré comme étant gagnant.

3.3 Evaluation des modèles

Une fois que la modélisation a été faite, il fallait calculer l'erreur relative de la manière suivante :

$$\text{erreur\_relative} = \frac{durée\ réelle\ cumulée\ sur\ le\ mois - durée\ prédite\ cumulée\ sur\ le\ mois}{durée\ réelle\ cumulée\ sur\ le\ mois}$$

On obtient le tableau suivant pour les 10 séries suivantes :

	SARIMA	SES	TES
IND MOB	+2.10%	-10.61	-16.06%
IND FIX	+21.14%	-12.94%	+14.68%
PRT MOB	-2.37%	+3.57%	-11.52%
PRT FIX	+2.24%	-9.82%	+6.84%
ITA MOB	-4.78%	-25.38%	+10.18%
ITA FIX	+5.75%	+90.58%	+4.98%
CHN MOB	-17.48%	-34.81%	-23.23%
CHN FIX	-22.72%	-33.87%	-19.95%
NLD MOB	+1.90%	+18.78%	+3.93%
NLD FIX	+3.16%	-17.63%	+2.87%
Nombre de +	6	3	6
Nombre de -	4	7	4
Nombre de win	7	4	7

Figure 14 : Tableau des erreurs relatives pour les 10 séries

Ce tableau des résultats présenté à la figure 10 présente une variable aléatoire **win** (gain du modèle) qui est vaut

- 1 si l'erreur relative est inférieure en valeur absolue à 15%
- 0 sinon

	Performance
SARIMA	70%
TES	70%
SES	40%

Figure 15: Performance des modèles pour les couples(country,carrier)

Le tableau ci-dessous présenté à la figure 11 fournit la performance des modèles le pourcentage du nombre de WIN sur l'ensemble des 10 séries:

3.4 Automatisation des tâches

Figure 16: Les différentes phases d'automatisation des tâches

L'architecture de la solution proposée est présentée dans le schéma de la Erreur ! Source du renvoi introuvable. ci-dessus.

Les étapes principales du processus sont les suivantes :

- Extraction des données provenant de la table EDH « c2r_2_0_calculation» au moyen d'un script en langage python exécuté
- Construction d'un dictionnaire contenant des triplets (pays, carrier, type d'accès) presents dans les données extraites dans l'étape précedente
- Pretraitement des donnés brutes en vue d'obtenir des series les series temporelles nécessaires à l'estimation des modèles predictifs

- Modélisation et présentation des résultats sous forme d'un tableau de prédictions. Ce tableau contiendra aussi des métriques permettant de surveiller le niveau de qualité des prédictions

Les performances de prédiction sur des triplets (country, carrier, access_type) pour les pays France et Portugal sont résumés dans le tableau ci-join

tableau_err_cumul_
mois_juillet_Portuga

Les colonnes suivantes sont présentes dans le tableau :

- Les colonnes A, B et C représentent les triplets (country, carrier, access_type)
- Les colonnes D et E représentent respectivement le nombre de jours d'entrainement et de test présents dans sur les triplets. La période d'entrainement va 1er Avril au 30 Juin 2021 et le mois de Juillet a été choisi comme base de test
- La colonne F représente la durée réelle du mois de juillet tandis que les colonnes G et H représentent respectivement la durée prédite par le modèle SARIMA et la durée prédite par le modèle TES sur le mois de Juillet.
- Les colonnes I et J représentent respectivement les erreurs relatives pour le modèle SARIMA et les erreurs relatives pour le modèle TES
- Les colonnes K et L représentent respectivement le WIN du modèle SARIMA et le WIN du modèle TES

La qualité de la prédiction est sensiblement dégradée par les données manquantes, soit dans la période d'estimation soit dans la période de prédiction. A noter que les données manquantes deviennent plus fréquentes quand la catégorisation se fait sur 3 variables (country, carrier, accès type), que dans le cas des résultats du tableau 1 où la catégorisation se fait sur 2 variables (country, access_type).

3.5 difficultés rencontrées

Dans l'ensemble, je n'ai pas rencontré trop de difficultés techniques. La notion de séries temporelles était quelque chose de nouveau pour moi. Dans un premier temps, j'ai suivi deux MOOC de coursera qui m'ont permis de bien assimiler la manipulation des séries temporelles ainsi que la modélisation des séries temporelles.

La seule difficulté majeure était de procéder à l'automatisation de la récupération des données. Cela est dû au fait que les données sont partitionnées sur Hue et qu'il fallait mettre la partition dans la condition where. À cause de cette difficulté, j'étais obligé de récupérer les données manuellement

Par la suite un collègue qui travaille dans l'équipe Big Data nous a aidé à surmonter cette difficulté.

4. Dimensions humaines et managériales

4.1 Weekly meeting

Un «weekly meeting» ou rencontre hebdomadaire a lieu chaque lundi dans lequel chaque membre de l'équipe y soumet son travail hebdomadaire sur Confluence. Notre manager prend 15 minutes pour parler des sujets généraux puis chaque membre parle de l'évolution de son sujet.

4.2 Conduite et organisation du projet

Date	Tâche réalisées
06/02/2021	Identification et compréhension du besoin de l'expert métier
09/02/2021	Présentation d'un cas d'usage des séries temporelles à l'expert métier
04/03/2021	Installation des ressources nécessaires pour accéder aux données
06/04/2021	Présentation du premier résultat d'exploration
07/05/2021	Présentation de l'exploration du carrier Orange IC
10/06/2021	Présentation de la performance des 3 modèles pour 5 pays choisis par l'expert métier
15/07/2021	Réunion pour discuter sur la possibilité d'automatiser les tâches
10/08/2021	Présentation finale des résultats
16/08/2021	Présentation des travaux à la personne en charge d'évoluer sur le projet
06/09/2021	Rapport d'exploration

Le tableau ci-dessus présente les différentes phases de la conduite de projet. Les différentes phases sont les suivantes :

- L'identification des besoins de l'expert métier dans lequel il fallait identifier et appréhender le problème de l'expert métier. Cette phase constitue l'étape primordiale. Il y a eu quelques changements concernant les couples à prendre à compte pour construire les modèles. Les différents couples étaient : le couple (pays, carrier) ou le couple (pays, type d'accès)
- Présentation du premier résultat d'exploration : cette première réunion faite le 06/04/2021 était l'occasion de présenter l'étude du carrier Vodafone. Par la suite Catherine et Sébastien(les experts métiers) nous ont orienté sur l'étude du carrier Orange IC que nous

avons présenté le 07/05/2021. Les résultats d'exploration se trouvent dans le fichier ci-joint :

exploration_carrier_orangeIC.pptm

- Présentation de la performance des 3 modèles pour 5 pays choisis par l'expert métier : suite à la présentation du carrier Orange, Cathérine et Sébastien nous ont fournis d'autres pays à explorer en vue de sélectionner par la suite les modèles à conserver dans notre étude. L'étude s'est fait en tenant en compte le couple(pays, type d'accès).Les résultats de cette exploration que nous avons présenté le 10/06/2021 sont disponible dans le fichier ci-joint :

pres_performance_3_modele.pptx

- Réunion pour discuter sur la possibilité d'automatiser les tâches : cette réunion,qui s'est déroulée le 15 /07/2021 a été organisée pour discuter sur une possible automatisation des scripts pour la mise en production.
- Présentation finale des résultats : qui est la dernière phase du projet et qui a eu lieu le 10/08/2021.

pres_resultats_finaux.pptx

- Rapport d'exploration : rapport pour expliquer ma démarche durant ce projet que j'ai soumis le 06/09/2021

4.3 Difficultés rencontrées

Il n'y a pas eu de difficultés majeures mais l'installation des drivers et outils nécessaires pour pouvoir accéder aux données a pris un certain temps. L'accès aux identifiants du VPN(Virtual Private Network) dans lequel je devais me connecter pour accéder aux données n'est possible qu'avec l'intervention des membres qui travaillent sur la sécurité.

5. CONCLUSION

Bilan

Ce sujet d'apprentissage, choisi par mon manager, m'a permis d'acquérir plusieurs compétences techniques et humaines.
Ce projet m'a permis de comprendre le métier d'un data scientist dans un environnement business. La première étape qu'un data scientist doit franchir avant d'entamer la partie technique est d'identifier le besoin de l'expert métier et comprendre le problème. J'ai pu acquérir des compétences sur une notion qui était nouvelle pour moi à savoir les séries temporelles, ce qui m'a permis de gagner en autonomie. J'ai pu mettre en œuvre les séries temporelles dans le cadre du projet en passant par la manipulation ainsi que la modélisation.
Ce projet m'a aussi permis d'acquérir des compétences humaines, à savoir travailler en équipe dans un environnement d'OBS/IT.

Evolution du projet

Ce projet sera poursuivi par un autre membre de l'équipe à qui j'ai eu à fournir tous les éléments nécessaires

Perspectives

Je vais continuer à approfondir mes connaissances sur les séries temporelles surtout les modèles que je n'ai pas eu l'occasion d'explorer.

Je compte approfondir aussi mes compétences en Hadoop et Spark, des notions que j'ai eu à faire à l'école mais qui n'ont pas été trop approfondies. Car des compétences en Data engineering sont nécessaires de nos jours pour être plus autonome en tant que data scientist.

6.Bibliographie

Éric Biernat, Michel Lutz. *Data Science: fondamentaux et étude de cas*. EYROLLES, 2015.

Fatoumata Dama, Christine Sinoquet. *Analysis and modeling to forecast in time series:*. Nantes, 2021.

Sadigov, Tural. *Practical time series analysis*. s.d. Février 2021. <https://d3c33hcgiwev3.cloudfront.net/_6f27156fa030a6e18372e06cf96d7b82_Week-1---slides--White.pdf?Expires=1632441600&Signature=Ig0IZHho~fcybT~aszj-7x7AI6cwBn88qZyxNY5Cy-YmYM7TrpkN3ud41PXSyJQrMOdDxkjBFX~qDjHazaYRpuIvaXlfF5Rn5L4whbBCMRQDa-pSixBpRILB7-8o7BvA>.

Annexes

Code informatique

- Fichier html présentant le code sur l'automatisation de la récupération des données

extraction_des_don
nees_hue.html

- Fichier pour l'automatisation des modèles ainsi que le tableau des résultats

series_temporelles_
et_tableau_resulats.