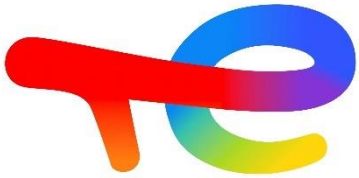


Rapport de stage

Dirigé par : Eurydice LAFFERAIRIE

Encadré par : Yoan SAINT-PIERRE

Réalisé par : Pierre BERNARD

Master 2 CSMI 2021 - Semestre printemps

Table des matières

Introduction

Contexte

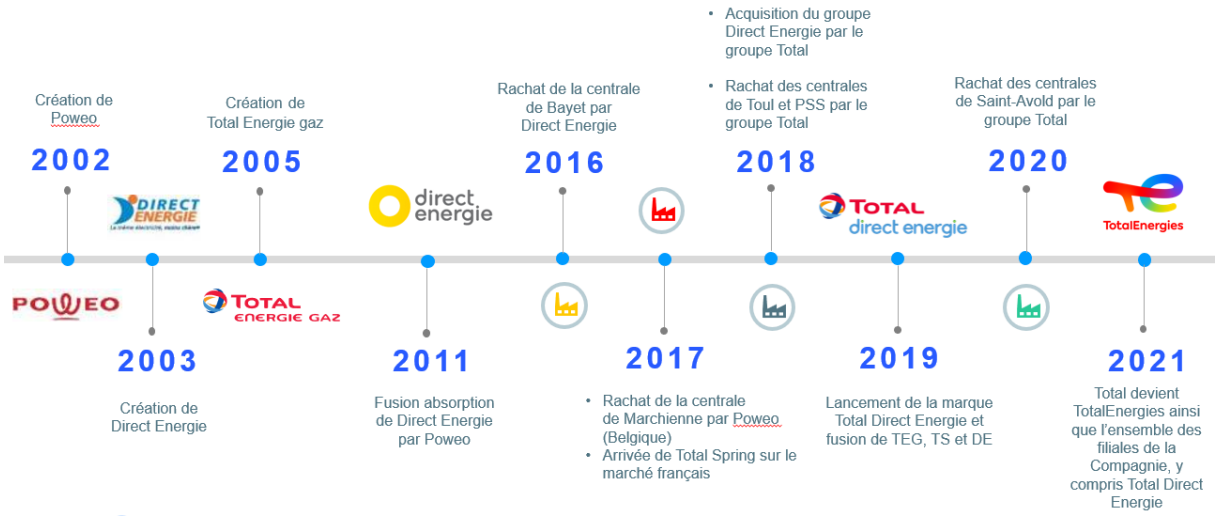
Le marché de l'énergie se divise en quatre catégories : la production, le transport, la distribution et la vente. Le 10 février 2000 [1], EDF perd le monopole de la production et de la vente d'électricité (afin d'assurer l'unicité du réseau de distribution, cette loi sur la fin du monopole électrique ne touche ni le transport ni la distribution).

EDF est alors obligé de vendre son électricité à un Tarif Règlementé de Vente (TRV) afin de permettre à des acteurs alternatifs de faire leur apparition en proposant des tarifs inférieurs au TRV [1] (exemple : apparition d'offres à -10% du TRV).

Bien qu'historiquement liés au groupe, la majorité des clients restent chez EDF, le marché finit par s'équilibrer et en 2019, une loi prévoyant la fin des TRV est votée le 30 juin 2023 [2].

Dans ce contexte, se démarquer de la concurrence devient de plus en plus compliqué et l'une des solutions, pour rester attractif et réduire la facture d'un client, est de l'accompagner et de l'aider à réduire sa consommation.

Histoire de la marque



En 2002 [3], Charles Beigbeder créé **Poweo**, société fournisseur/producteur d'électricité et de gaz. Un an plus tard, par la création de **Direct Energie** par Xavier Caïtucoli, Direct Energie rachète et absorbe Poweo en 2011. Cette fusion sert de tremplin au groupe qui prend de l'ampleur jusqu'en 2018 [3] et atteint suffisamment de part de marché pour attirer l'attention de la multinationale **Total** dirigée par Patrick Pouyanné qui la rachète alors.

L'année suivante, Total fusionne ses filiales **Total Energie Gaz** (fournisseur/producteur de gaz fondé en 2005 [3]), **Total Spring** (fournisseur/producteur d'électricité fondé en 2017 [3]) et **Direct Energie** sous une même marque : **Total Direct Energie.**

Finalement, afin de se défaire de son image de groupe producteur de pétrole et d'incarner un groupe multi-énergies, le groupe Total et l'ensemble de ses filiales est renommé en mai 2021 [3] **TotalEnergies**. Afin de désigner la branche Total Direct Energie au sein de Total, on parle désormais de **TotalEnergies PGE – France** (Power&Gaz Europe).

Place au sein du groupe

Le groupe Total Energies est composé de plus de 100 000 collaborateurs partout dans le monde, qu'on peut diviser en 5 grands secteurs d'activité :

- Exploration et Production : recherche de pétrole/gaz naturel
- Raffinage et Chimie : transformation du pétrole brut et du gaz naturel en produits finis
- Marketing & Services : activités liées aux stations-services
- Trading & Shipping : négociations et transport des produits
- **Gaz, Energies Renouvelables & Electricité** :
    - Gaz
    - Energies Renouvelables
    - Neutralité Carbone
    - Finances
    - Stratégie, Croissance & Ressources Humaines
    - **Electricité et Gaz Europe : Power & Gas Europe (PGE)** dans 6 pays européens
        - **PGE – France**

Ce stage se déroule au sein de PGE – France, fournisseur et producteur d'électricité et de gaz en France au sein du groupe Total Energies.

Les Mission de PGE – France sont la commercialisation de gaz et d'électricité, les services associés aux clients particuliers (aussi appelés clients BtoC), aux clients professionnels (aussi appelés clients BtoB) et le développement, la gestion et l'optimisation des centrales au gaz.

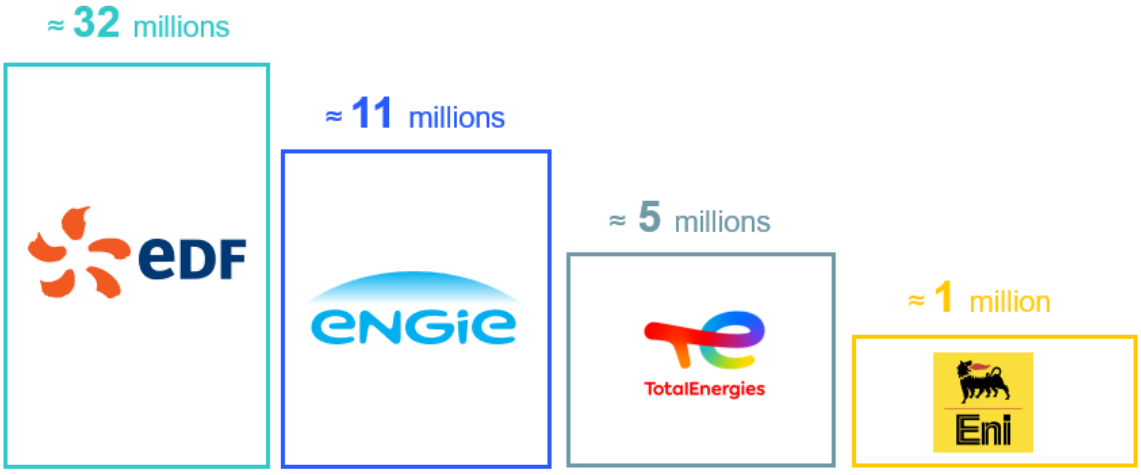


Figure 1 - Place de TE sur le marché de l'électricité et du gaz [4]

Avec ses 5 millions de sites et clients (dont 80% de particuliers), PGE – France, en tant que 3ème fournisseur d'électricité/de gaz en France, s'inscrit tant comme producteur bas carbone que comme fournisseur responsable, en tant qu'acteur majeur dans l'ambition de Total Energies d'atteindre la neutralité carbone en 2050 et d'ici 2030, une place dans le top 5 des producteurs d'énergies renouvelables (solaire/éolien).

Objectifs

Figure 2 - Compteur linky

Les compteurs d'électricité intelligents Linky produisent des données qui peuvent aller jusqu'à la maille seconde. Cet énorme volume de données est encore peu exploité.

Ainsi, l'objectif principal de ce stage est de donner de la valeur à ces données en extrayant des informations utiles qui nous permettront de guider le consommateur vers des économies d'électricité ou de gaz.

La masse de données que possède Total Energies ne concerne pas seulement directement la consommation de ses clients. En effet, chaque interaction avec le client est enregistrée et ajoutée à une base de données. Ces informations sont précieuses pour les autres corps de métier (Service Après Ventes, par exemple) qui ont tout intérêt à être capables de prédire avec précision ces interactions, afin de manager au mieux leur effectif. Ainsi, j'ai également pu travailler sur la prédiction des appels téléphoniques sur plusieurs canaux de communication de Total Energies.

Contraintes

Le compteur intelligent Linky peut, dans le cas des clients qui possèdent la clef conso-Live, décomposer leur consommation jusqu'à la maille seconde. Ainsi, avec 5 millions de clients, les volumes de données sont donc considérables. Il devient alors impossible d'utiliser le framework habituel de la DataScience, Pandas/Python. Ainsi, du moins tant qu'on n'a pas encore agrégé les données à une maille plus grosse, il est nécessaire de travailler dans un autre framework de gestion calcul de base de données distribué : Spark ou son API python PySpark.

Framework

Spark – PySpark

Spark (ou Apache Spark) est un framework open source de calcul parallèle [5]. Il s'agit d'un ensemble d'outils et de composants logiciels structurés selon une architecture définie. Ce produit est un cadre applicatif de traitements big data pour effectuer des analyses complexes à grande échelle. Son

utilisation est donc essentielle pour les étapes de pré-processing des algorithmes utilisés au cours de ce stage.

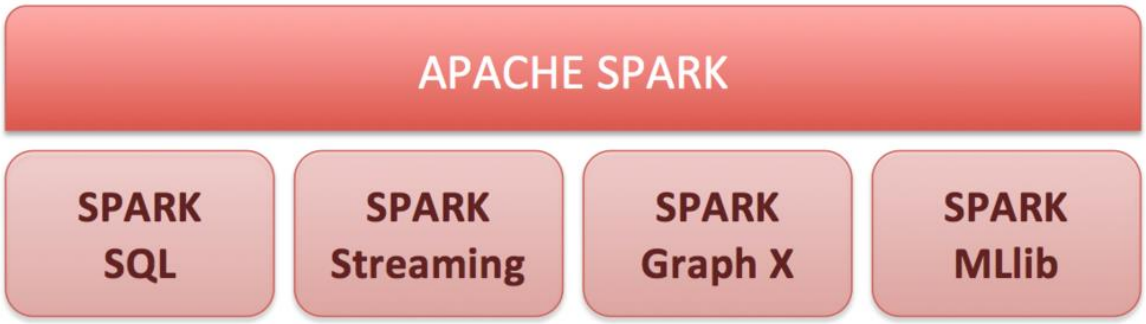


Figure 3 - Outils développés par Apache Spark

En particulier, Spark SQL permet d'exécuter des requêtes en langage SQL pour charger, manipuler et transformer des données [5]. Si le langage SQL est utilisé pour traiter des bases de données relationnelles, dans Spark, il permet de traiter n'importe quelles données, quel que soit leur format d'origine [5].

PySpark est l'API Python de Spark [6], et permet l'utilisation de ce cadre de calcul parallèle tout en gardant la simplicité d'écriture du langage Python. Cette API est très populaire et bénéficie d'une forte communauté active ainsi que d'une bonne documentation.

Databricks

A l'instar de Colaboratory de Google, Databricks est une plateforme web de partage de notebooks créée par les créateurs d'Apache Spark [7]. La plateforme est conçue autour de Spark et propose un outil de création et de management de clusters de calcul de taille conséquente (celui utilisé au cours de ce stage était un cluster de 28 GB de RAM / 8cores) en faisant ainsi un Framework de travail idéal pour le Big Data [7].

Outre la simplicité du travail en équipe qu'apporte la possibilité de travailler à plusieurs sur le même notebook, le véritable point fort de Databricks est la facilité avec laquelle il est possible de passer un projet de l'étape de développement à celle de mise en production.

En effet, après avoir écrit un notebook, il est possible de programmer le run de ce notebook à intervalle régulier. Il est possible également de programmer des notifications mails en cas de réussite/échec du run de ce notebook afin de tracer les résultats des algorithmes au cours du temps.

Enfin, le jumelage du compte Databricks à un repository GitHub permet l'utilisation des outils d'intégration continue et le confort de bénéficier d'un historique mis à jour automatiquement.

Prédiction d'appels

Les managers des équipes de contact clients ont un délai légal de 3 mois pour signer de nouveaux contrats, former les nouveaux membres de leurs équipes ou réajuster leur effectif. Afin d'éviter une saturation du service d'appel ou au contraire, un nombre d'appels insuffisant par rapport à la taille de l'équipe, il est essentiel de réussir à fournir tous les mois une prédiction du nombre d'appels à la maille hebdomadaire des 3 mois suivants.

Les différents services en contact avec la clientèle (Service Client particulier, Vente particulier…) sont soumis à des afflux d'appels dont les propriétés (saisonnalité/trend) peuvent parfois beaucoup différer. Aussi, il faut créer un modèle par service/par flux d'appel (numéro « ClickToCall » disponible en ligne, ligne 3099, …).

Prophet

Lorsque l'on parle de série temporelle, il est important de définir les notions de trend et de saisonnalité.

- La trend (tendance) est l'orientation générale de la série temporelle sur une période assez longue.
- La saisonnalité de la série temporelle est sa périodicité, son motif lorsqu'on lui retire sa tendance.

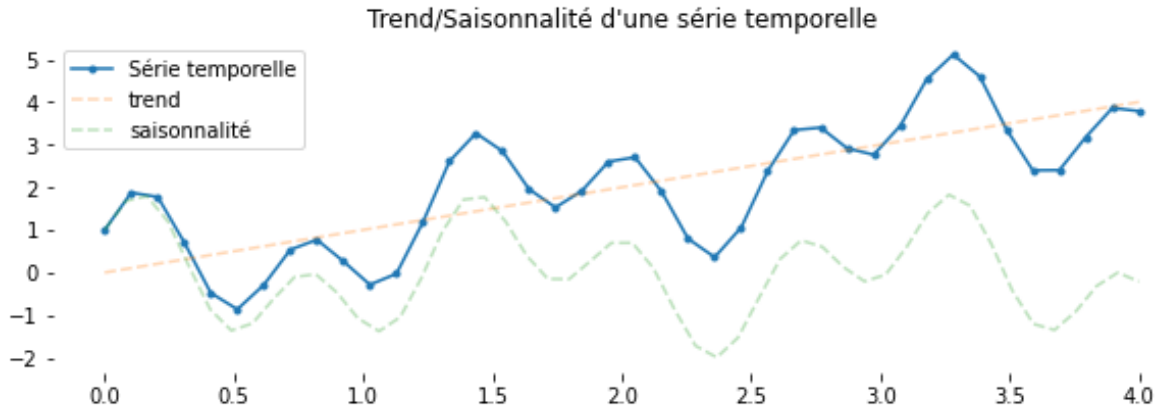


Figure 4 – exemple de décomposition Trend / Saisonnalité manufacturé

Cette décomposition, cruciale pour la compréhension de la série temporelle, est faite automatiquement par un module python de prédiction de série temporelle mis au point par Facebook : Prophet.

Le modèle fourni par Prophet considère chaque série temporelle comme une somme d'une fonction décrivant la saisonnalité, la tendance et les effets de vacances/jour fériés/jours extraordinaires de la série temporelle [8].

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

Figure 4 - Décomposition trend/season/holiday

Où :

- **y(t)** est l'output qu'on cherche à prédire
- **g(t)** est une fonction représentant le trend de la série temporelle.

Prophet fournit l'option d'approcher ce trend comme linéaire par morceaux ou logistique par morceaux [8].

$$g(t) = \frac{C(t)}{1 + x^{-k(t-m)}}$$

Figure 5 - Formule d'une fonction logistique

Le découpage de g(t) en morceaux linéaires ou logistiques est fait automatiquement par Prophet de manière à minimiser la différence entre la série temporelle et la prédiction pour un nombre de points de changement de trend (changepoint) donné.

- **s(t)** est une fonction qui décrit les motifs saisonniers de notre série temporelle.

  Ces motifs peuvent être prédits à la maille de la semaine (pas d'appels le dimanche puisque le service est fermé par exemple) ou à la maille annuelle (davantage d'appels dans les périodes de déménagements, vers septembre par exemple). Dans notre cas, le métier s'intéresse à des prédictions de l'ordre du nombre d'appels par semaine, aussi nous n'utiliserons que la saisonnalité annuelle.

  La prédiction de la saisonnalité faite par Prophet repose sur une décomposition en série de Fourier [8].

$$s(t) = \sum_{n=1}^{N} (a_n \cos(\frac{2\pi nt}{P}) + b_n \sin(\frac{2\pi nt}{P}))$$

Figure 6 - Décomposition en série de Fourier

- **h(t)** est une fonction qui décrit les effets de vacance / jour férié / évènement exceptionnel comme confinement soudain lié au Covid-19 auquel notre modèle est susceptible d'être sensible. D'après un retour d'expérience métier, le nombre d'appels est peu sensible aux vacances, aussi seuls les jours fériés sont utilisés lors de nos prédictions.
- **ε(t)** représente le résidus de notre prédiction.

Prophet fournit également la possibilité d'ajouter des régresseurs personnalisés qui pourraient sembler importants à la prédiction, comme le budget investi dans la publicité mois par mois.

Les avantages de Prophet comparés à une ARIMA (AutoRégression Integrated Moving Average) plus traditionnelle sont [9]:

- La précision et la vitesse de fitting (fit fait en Stan)
- La facilité d'utilisation grâce à de nombreuses fonctions automatisées
- Beaucoup de paramètres à tuner pour améliorer la prédiction initiale
- L'interprétabilité des résultats (possibilité d'afficher la décomposition trend/saisonnalité/holidays pour comprendre l'output du modèle)

Neural Prophet

Si les prédictions de Prophet sont très satisfaisantes la plupart du temps, dans le cas de séries temporelles à croissance exponentielle, un modèle à trend linéaire ou logistique par morceaux sous-performe [10]. Aussi, la bibliothèque Neural Prophet a été créée afin de bénéficier de la simplicité d'utilisation de l'API de Prophet tout en utilisant un Réseau de Neurone Autorégressif (AR-Net) pour mieux décrire les cas non-linéaires.

On parle d'un réseau de neurone autorégressif lorsque l'output est directement obtenu de plusieurs temps précédents de l'input. Contrairement à un Réseau de Neurone Récurrent (RNN) où les temps précédents sont fournis via des états cachés.

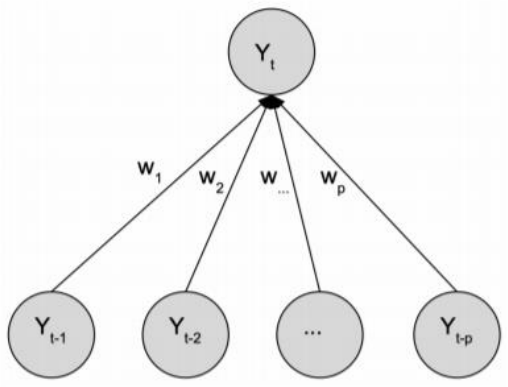


Figure 5 - Réseau de Neurone Autorégressif

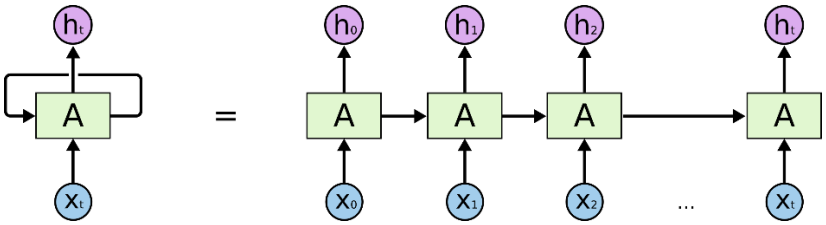


Figure 6 - Réseau de Neurone Récurrent

Prophet était déjà utilisé pour des prédictions d'appels avant mon arrivée en stage. Aussi, l'une des premières missions de mon stage a été de tester la bibliothèque Neural Prophet et de comparer ses performances avec Prophet.

Prophet vs Neural Prophet

Afin d'étudier les performances de Neural Prophet, nous avons sélectionné la série temporelle avec le trend le plus exponentiel que nous avions à notre disposition et dont il existait déjà une prédiction basée sur un modèle Prophet.

L'objectif de ce premier projet était de me familiariser avec l'utilisation de Prophet, Databricks, et potentiellement de découvrir un cas d'usage où l'utilisation de Neural Prophet serait préférable à celle de Prophet.

Après avoir créé un modèle Neural Prophet, l'avoir tuné à l'aide de grids searchs en utilisant la MAE des cross validations pour sélectionner les meilleurs hyperparamètres, on obtient les prédictions suivantes :

Figure 7 - Comparaison Prophet vs Neural Prophet

Si les deux modèles ont tous les deux réussi à prédire le creux très récurrent de début d'année en 2021, Neural Prophet fait ici de l'overfitting : la MAE sur la zone de validation est bien supérieure (MAE relative au nombre d'appels maximum : 15.5 %) à celle de la zone de train qui était particulièrement basse.

Prophet en revanche prédit la série temporelle de manière très satisfaisante, la zone de validation ayant une MAE comparable à celle de la zone de train. (MAE relative au nombre d'appel maximum : 5.71%).

Si la série temporelle choisie pour comparer Prophet à Neural Prophet semblait de prime abord avoir un fort trend, finalement elle était probablement trop linéaire pour justifier l'utilisation de Neural Prophet.

Aussi, par la suite, nous continuerons d'utiliser le modèle habituel, Prophet.

Prédiction d'appel – Méthodologie

Au cours de ce stage, j'ai eu l'occasion de travailler sur plusieurs projets de prédiction d'appels. Pour chacun d'entre eux, nous avons utilisé la boucle de communication suivante :

1. Discussion initiale avec le métier pour comprendre la nature des données à notre disposition et les attentes sur la prédiction à fournir
2. Création d'un modèle de prédictions naïf
3. Discussion avec le métier des anomalies de prédictions repérées et de jeux de données exogènes potentiellement corrélés à notre output
4. Mise à jour du modèle en tenant compte de l'expertise métier
5. Retour en étape 3. Si les résultats ne sont pas satisfaisants, passage à l'étape suivante, sinon :
6. Comparaison des résultats aux prédictions déjà existantes

Prédiction d'appels – Service Client Pro

Description du jeu de données et de l'output attendu

Pour cette prédiction, on dispose du nombre d'appels journaliers des clients professionnels.

Afin de pouvoir avoir le temps de prendre en compte les prédictions lors de la formation des équipes de standardistes, les managers ont besoin tous les mois d'une prédiction sur 3 mois des futurs appels à la maille hebdomadaire. Dans le cas de semaines à cheval sur deux mois, deux prédictions sont attendues pour cette semaine : le prorata de la semaine présente dans le premier mois, et le prorata de la semaine présente dans le second mois.

Aussi, en amont, il est nécessaire d'effectuer un travail de resampling pour passer de la maille journalière à la maille hebdomadaire, la maille de la prédiction.

En aval, il faut faire une prédiction supplémentaire : il faut prédire l'importance de chaque jour de la semaine sur le nombre total d'appels. On peut ainsi générer les proratas mentionnés plus haut et remplacer les prédictions des semaines à cheval sur deux mois par ces proratas. Cette prédiction est faite à partir du modèle LightGBM (une version plus légère en calculs de XGBoost).

Création du modèle

Si un modèle naïf permet déjà d'obtenir une première prédiction raisonnable, on remarque tout de même l'apparition, en début 2017 et fin 2018, de deux pics d'appels particulièrement hauts qui n'ont pas été bien prédits par le modèle.

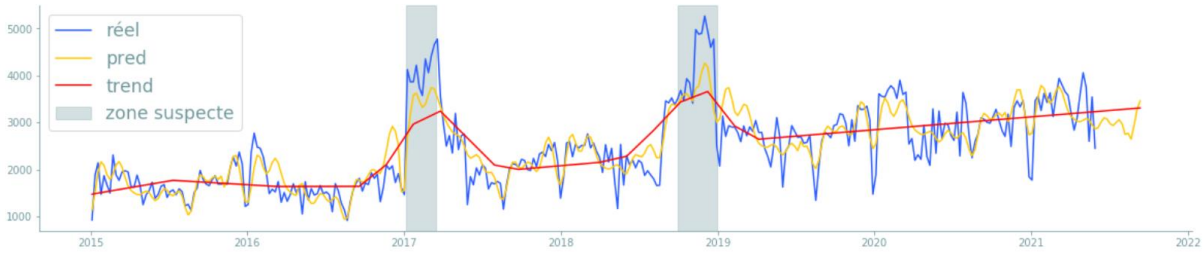
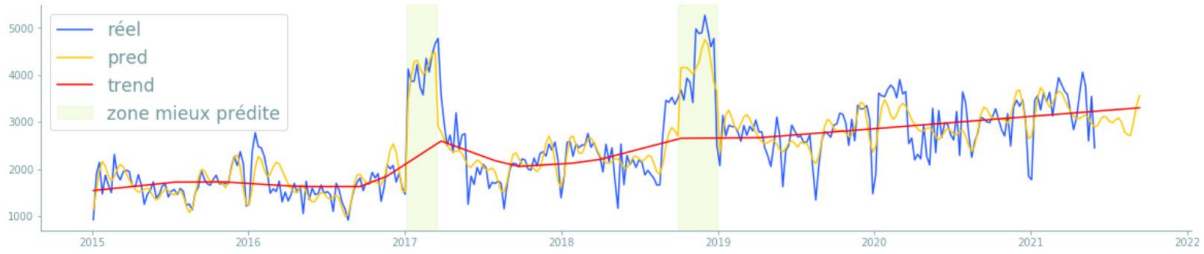


Figure 8 - Prophet PRO Baseline Model (pics suspects)

Après discussion avec le métier, ces pics d'appels sont encore à ce jour inexpliqués. Par conséquent, afin d'éviter que ces pics n'impactent la saisonnalité de notre modèle, on introduit une variable exogène artificielle, une colonne flag binaire qui indique au modèle qu'on se trouve dans l'une des périodes suspectes.



De la même façon, on flag aussi la période du premier confinement qui a réduit la plupart de nos flux d'appels.

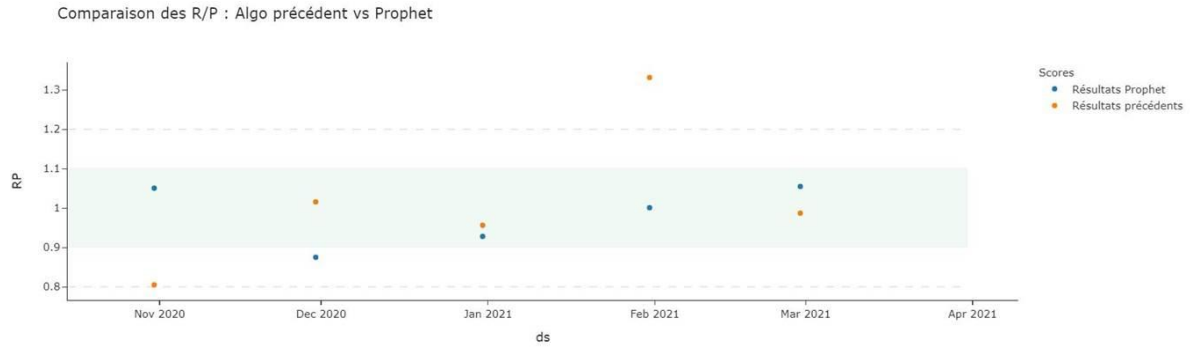
Comparaison des résultats

Afin de valider les résultats des prédictions, le métier utilise une métrique appelée le R/P (Réel/Prédit). Le tunnel standard de qualité attendu par la prédiction actuelle est de 90%-110%.

Aussi, afin de s'assurer que notre modèle propose des prédictions de meilleure qualité que ce qui est déjà fait, on compare le R/P mois par mois de notre modèle à d'anciennes prédictions enregistrées. Pour ce faire, on ne train notre modèle que sur l'historique d'appels dont l'ancien modèle disposait, et on reproduit la prédiction qu'il aurait pu faire en lui fittant mois par mois les données qui apparaissent.

Figure 9 - Prédiction d'appels PRO - prédiction mois par mois

Ici, les pointillés verticaux représentent les jonctions successives entre train et prédiction de notre modèle en début de mois. A partir de cette prédiction, on peut alors calculer nos R/P mois par mois et les comparer aux données fournies par le métier.



4 mois prédits sur les 5 sont dans le tunnel de qualité RP 90%-110% (contre 3 du côté de la prédiction actuelle). Et en inspectant le mois qu'on a surévalué (décembre 2020) de plus près et après discussion avec le métier, on remarque en particulier que la surévaluation serait dûe à une semaine de confinement mal interprétée par notre modèle.

On peut aussi calculer la MAE de chaque modèle sur cette période :

- MAE Prophet : 851.03 (erreur relative au maximum d'appel : 17.02%)
- MAE Ancien Modèle : 1613.13 (erreur relative au maximum d'appel : 32.26%)

Ces résultats ayant été jugés très satisfaisants, le projet a été passé à l'équipe Big-Data, responsable de la mise en production.

Prédiction d'appel – Service Client Vente Particulier

Description du jeu de données, de l'output attendu et des enjeux

Le Service client responsable de la gestion des appels liés à des nouveaux contrats (appelés Vente) reçoit des appels de différents flux en fonction de leur origine. Pour des raisons de changement récents d'architecture de la redirection de certains flux du Service, seuls les flux 3099 (numéro présent sur les campagnes publicitaires) et le CTC (Click To Call, numéro disponible en ligne près des campagnes publicitaires) furent jugés susceptibles d'aboutir à une prédiction correcte.

Les prédictions étaient faites jusqu'alors à la main chaque début de mois en utilisant une fonction auto native d'Excel : Holt-Winter. Ces prédictions étaient ensuite envoyées aux managers qui pouvaient les réévaluer en fonction de leur expérience métier. L'un des principaux enjeux de ce projet était donc d'automatiser cette prédiction et si possible de l'améliorer.

Les données concernant ces flux d'appels sont elles aussi à la maille jour et les contraintes sur l'output sont les mêmes que pour la prédiction précédente, aussi on utilisera la même méthodologie en pré-processing et post-processing que mentionné précédemment.

Dans le cas des nouveaux contrats, il est également intéressant d'introduire le budget prévisionnel mensuel alloué à l'ensemble des publicités TotalEnergies sous la forme d'une variable exogène. Il a donc été nécessaire de réaliser une étape de pré-processing supplémentaire à l'aide de requêtes SQL pour aller chercher ces informations sur le compte Google Analytics de l'entreprise.

Flux d'appels CTC

Comme précédemment, afin de valider nos prédictions, on compare nos R/P à des mois dont le métier possédait encore le R/P de leurs prédictions.

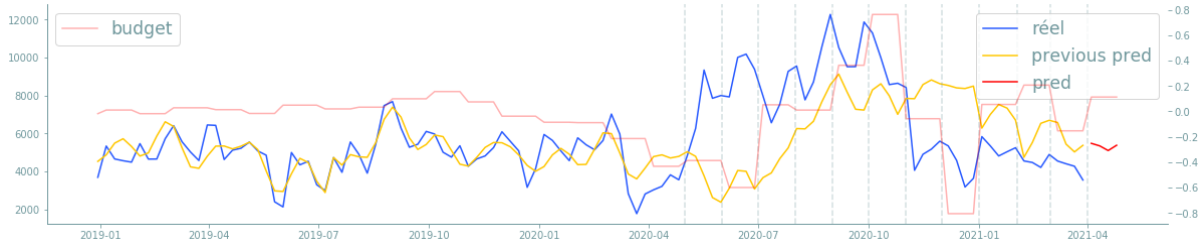


Figure 10 - Prédiction CTC

Sur la figure ci-dessus, les pointillés verticaux sont les jonctions successives entre train et prédiction. La démarcation orange/rouge représente le dernier mois de prédiction.

Contrairement aux autres prédictions vues précédemment, la série temporelle décrivant le flux CTC n'a que très peu de saisonnalité et a considérablement changé à partir de mai 2020. Dans ces conditions, le modèle a du mal à saisir à décrire les variations du flux d'appels.

Le modèle précédent avait pourtant mieux fonctionné :

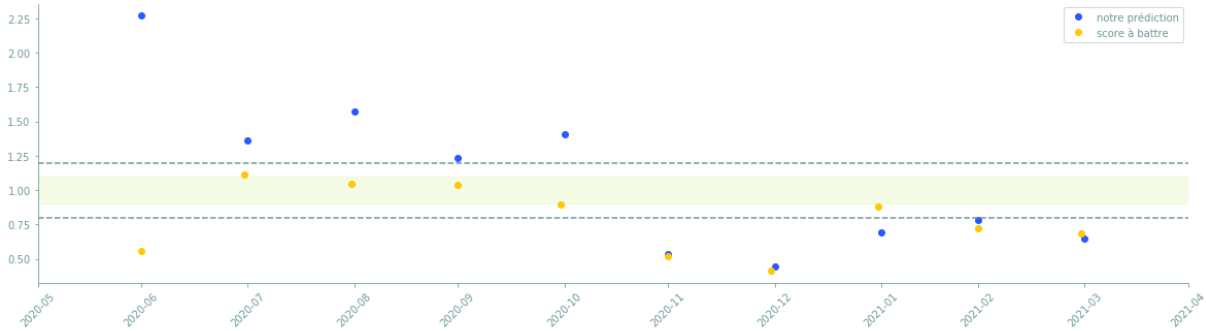


Figure 11 - Comparaison R/P CTC

On constate cette différence de qualité dans la prédiction aussi du côté des MAE :

- MAE Prophet : 14 160.55 (soit une erreur relative au maximum de 118%)
- MAE Ancien modèle : 4 992,27 (soit une erreur relative au maximum de 41.7%)

Après discussion avec le métier, cette différence de qualité dans la prédiction des appels s'explique par le fait que le R/P était calculé après réajustement humain par les managers. Aussi, l'idée d'aller plus loin avec ce flux d'appel a été laissée de côté pour l'instant.

Flux d'appels 3099

Encore une fois, après avoir créé un modèle baseline enrichi ensuite par plusieurs discussions avec le métier, on prédit le flux d'appels mois par mois pour le 3099 sur les mois dont on dispose du R/P de référence afin de valider nos prédictions.

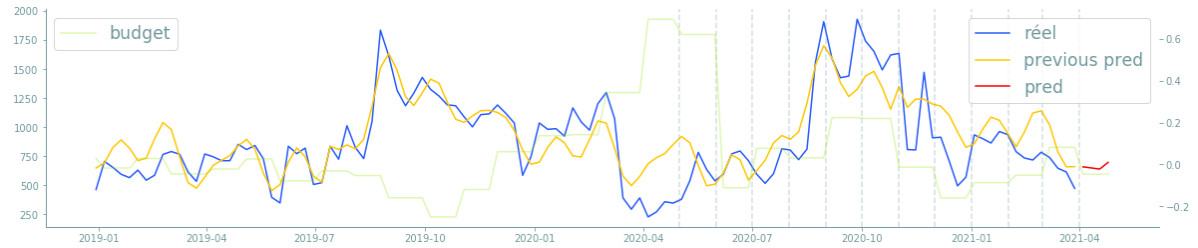


Figure 12 - Prédiction 3099

On arrive cette fois à une prédiction bien plus raisonnable qui réussit à prédire la bosse saisonnière de début d'année (voir janvier 2020, janvier 2021) notamment.

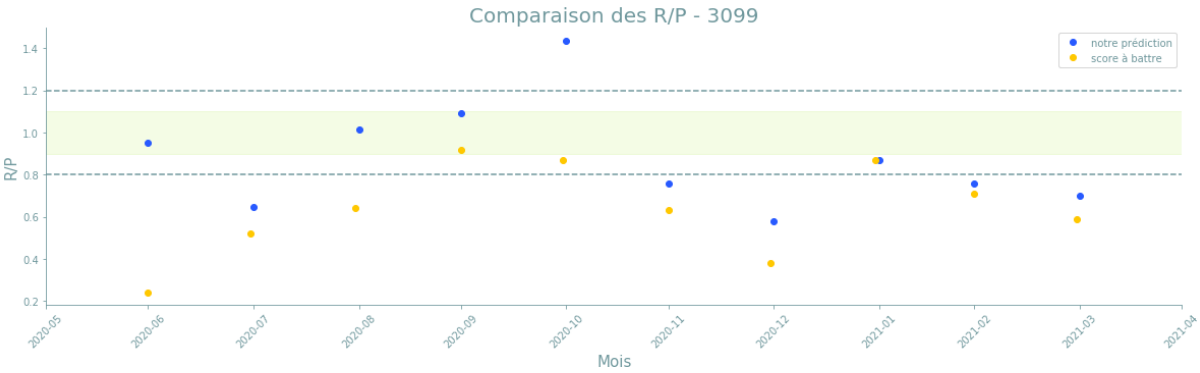


Figure 13 - Comparaison R/P 3099

Note : Comme dans le cas du CTC, les résultats ci-dessus concernant les R/P de l'ancien algorithme étaient calculés après réévaluation humaine par le manager avec son expérience du flux d'appel.

Si nos prédictions sont rarement exactement dans le tunnel de précision standard (90-110%) évoqué lors des prédictions précédentes, elles sont tout de même de bien meilleure qualité que les prédictions précédentes.

On peut constater cette différence aussi du côté des MAE :

- Ancienne MAE = 2477.80 (erreur relative au maximum de 120%)
- Nouvelle MAE = 1064.6 (erreur relative au maximum de 50%)

Aussi, les résultats ayant été très bien reçus par le métier, le modèle a été passé à l'équipe Big-Data, et est lui aussi passé en mise en production.

Suivi Conso

Bilan Hiver

Un des engagements de TotalEnergies est d'accompagner ses clients pour mieux comprendre leur maitrise d'énergie à l'aide d'outils de monitoring, de conseils de campagnes mails.

L'objectif du mail « bilan hiver » est d'estimer la consommation de chauffage de chaque client et de les aider à se repérer par rapport aux autres clients de son département. Ainsi, les clients peuvent recevoir un mail du type « cet hiver nous estimons ta consommation de chauffage à 90% de ta consommation totale, c'est beaucoup comparé aux autres clients de ton département ».

Pour faire cette estimation, on sépare les clients en deux groupes, pour lesquels on appliquera deux méthodologies différentes :

- **Les clients chauffés à l'électricité** qui disposent de beaucoup de données grâce aux compteurs Linky qui enregistrent les données de manière très fine (maille jour ou seconde s'ils disposent de la clef conso-live).

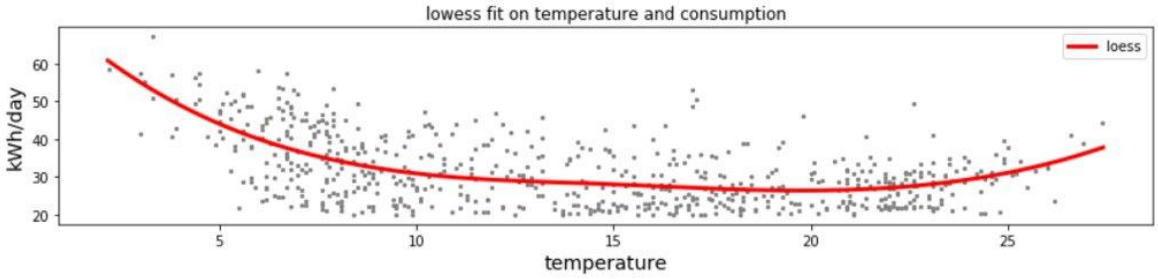


Figure 14 - Exemple de client chauffé à l'électricité

Pour chaque client, on étudie l'année en cours et on fait une régression locale (ou LOcally Estimated Scatterplot Smoothing LOESS [11] en anglais) à partir de laquelle on calcule la pente. En étudiant ensuite cette pente, on peut séparer les données en deux jeux : celles qui ont une pente négative parce que le client chauffe son domicile, considérées comme faisant partie de l'hiver, celles ayant une pente positive parce que le client climatise son domicile et les autres.

On peut alors interpréter la consommation médiane dans la période à faible pente comme étant la consommation d'énergie indépendante de la température, la consommation hors chauffage d'un client (chauffe-eau, cuisson, appareils électro-ménagers, …).

En faisant l'hypothèse assez lourde que la consommation hors chauffage d'un client est peu dépendante du moment de l'année, on peut alors soustraire cette consommation hors chauffage médiane pour obtenir une approximation de la consommation de chauffage d'un client chauffé à l'électricité.

Pour ce type de client, l'estimation du chauffage avait déjà été faite avant mon arrivée en stage. J'ai pu en revanche implémenter une version PySpark (et donc parallélisé) du calcul de la LOESS qui était extrêmement coûteux en version Python et qui sera probablement utilisé dans plusieurs autres futurs projets.

- **Les clients chauffés au gaz** qui disposent de moins d'historique car les relèves de gaz sont mensuelles.

  Pour ces clients, il a fallu faire preuve de créativité pour appliquer une technique similaire avec moins de données (le calcul de la LOESS n'étant plus possible).
  Plutôt que d'utiliser la pente globale du nuage de points pour faire la délimitation hiver/autre (le cas des clients climatisés n'étant plus problématique lors de l'étude de la consommation de gaz), pour chaque client, on trace la frontière au niveau de sa médiane des températures.

  A partir de là, la méthodologie redevient la même : on fait l'hypothèse forte que le client garde une consommation hors chauffage relativement constante par rapport à la période de l'année et on calcule la consommation hors chauffage.

  On soustrait ensuite cette quantité à la consommation du client enregistrée au dernier hiver pour obtenir une estimation de sa quantité de chauffage.

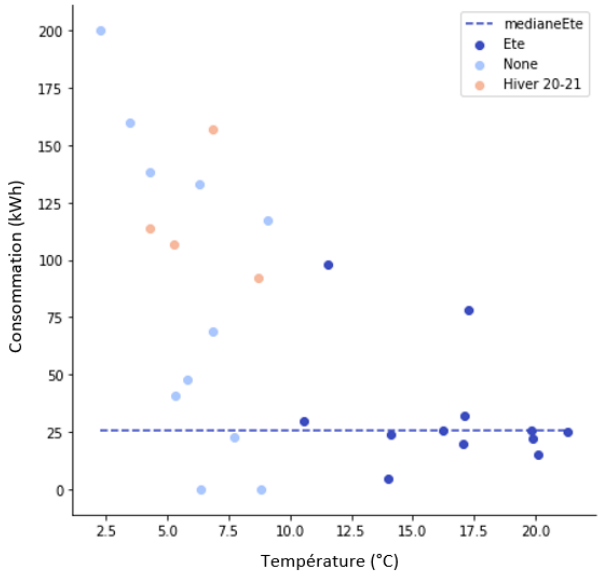


Figure 15 - Exemple de client chauffé au gaz

Ensuite, pour l'électricité comme pour le gaz, une fois la quantité de chauffage consommé estimée, on calcule la proportion de la consommation qu'on attribue au chauffage sur tout l'hiver.

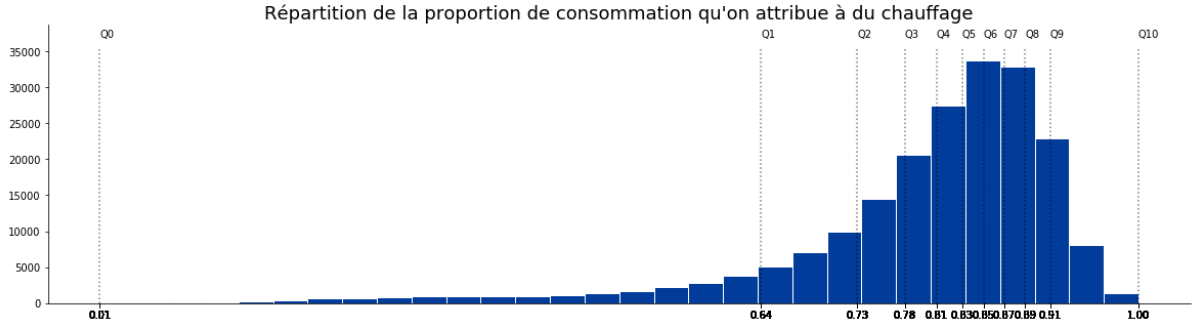


Figure 16 - Répartition des estimations de chauffage

Dans le cas électricité comme dans le cas gaz, on avait fait une hypothèse forte sur la consommation du client : on suppose qu'il consomme autant d'électricité/de gaz hors chauffage, peu importe la période de l'année. Afin d'éviter d'envoyer un mail à des clients dont le style de vie serait trop éloigné de cette hypothèse, on envoie cette campagne uniquement aux clients ayant un fort coefficient de corrélation entre température et corrélation.

De même, afin d'éviter d'envoyer des valeurs trop extrêmes* issues d'estimations aux clients, on filtre les clients à droite et à gauche de la courbe ci-dessus.

Une fois les différents filtres appliqués, on peut alors diviser chaque département en 3 groupes :

- Ceux qui chauffent le moins (les 25% à gauche restants)
- Ceux qui chauffent le plus (les 10% à droite restants)
- Les autres

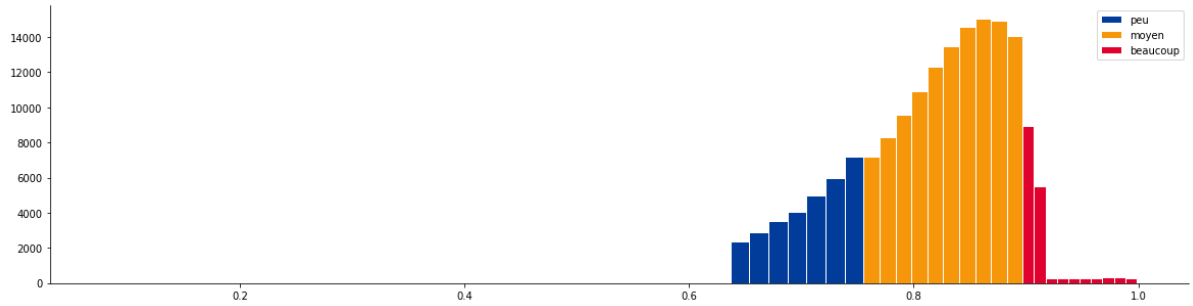


Figure 17 - Répartition % de chauffage par groupe

(*) Note : Sur la figure précédente, on voit qu'on a quand même envoyé le mail à quelques clients ayant entre 95% et 100% de chauffage détecté alors que plus haut, on avait mentionné un filtre évitant les valeurs trop extrêmes. Dans le cas des clients chauffés au gaz qui n'utilisent leur gaz que pour le chauffe-eau et ont un mode de cuisson électrique, il est naturel de considérer que leur consommation hors-chauffage est globalement constante avec la période de l'année.

Si ce projet n'a pas été mis en production parce que son besoin est bien plus ponctuel que la prédiction d'appels, il a permis l'envoi d'un mail « bilan hiver » aux clients concernés et sera probablement réutilisé en début d'année prochaine.

Analyse d'impact

L'analyse de l'impact d'un traitement médicamenteux, d'une offre, d'un conseil, d'une campagne ou d'un service peut avoir plusieurs intérêts [12] :

- Cibler plus précisément la population avec laquelle une offre va mieux fonctionner. Dans le cas d'une offre qui coûterait de l'argent à une entreprise, notamment.
- Vérifier l'efficacité d'un service proposé
- Utiliser cette analyse comme argument lors de démarchage ou de campagnes publicitaires
- Voir même dans le meilleur des cas, acquérir un label officiel de réduction d'énergie

Pour ces deux dernières raisons notamment, le service marketing, sous les conseils de mon encadrant qui avait un background d'économétrie, s'est intéressé à la possibilité de produire des analyses d'impact.

Pour me familiariser à l'utilisation d'une librairie d'économétrie dédiée à l'analyse d'impact, EconML, j'ai eu pour mission d'analyser dans un premier temps l'impact de la clef conso-live qui permet de

décomposer la facture d'électricité à la maille seconde et permet par la suite d'obtenir de nombreux conseils (détection des appareils en veille électrique, étude du moment de plus grosse consommation de la journée, décomposition des courbes de charges afin de détecter différents appareils électroménagers, …) issus d'algorithmes se basant sur ces données. Ce service est considéré comme ayant un impact probablement très fort sur la réduction de la consommation d'énergie les premiers mois suivant l'utilisation de la clef donc l'impact serait facilement détectable par une analyse d'impact.

Terminologie

Il convient de définir certains termes avant d'aller plus loin [13]:

- Le traitement est l'input (ex : l'utilisation du service ou non, la réduction proposée par une offre, le prix d'un produit, la quantité d'un médicament administré à un malade …). Il peut être binaire ou continu. Dans notre cas, on se situe dans l'utilisation d'un traitement binaire (dispose d'une clef conso-live ou pas).
- L'outcome est la variable d'intérêt (ex : la quantité achetée, la réponse au traitement médicamenteux …). Dans notre cas, on s'intéresse à la réduction de la consommation d'énergie après l'acquisition de la clef conso-live.
- Le (Conditionnal) Average Treatment Effect (CATE/ATE) représente l'effet moyen du traitement sur l'ensemble de la population ou sur un échantillon ciblé (l'impact du traitement est-il plus important chez les jeunes ?).
- Les variables instrumentales sont des variables qui impactent **uniquement** le traitement **mais pas** l'outcome directement.
- Les variables de contrôle (aussi appelés covariables) impactent le traitement **et** l'outcome (dans notre cas, la variance de la température dans une région donnée, par exemple).
- Les features sont des variables de contrôles auxquelles on s'intéresse particulièrement pour étudier l'impact dans un échantillon de la population (dans la question « l'impact du traitement est-il plus important chez les jeunes ?» par exemple, l'âge est une feature).

Formulation mathématique du problème

Soit $Y(\mathrm{t})$ une variable aléatoire qui correspond à la valeur de l'outcome si on soumettait un échantillon à un traitement $\mathrm{t}\in\mathrm{T}$. Etant donné deux vecteurs de traitements t_0, $\mathrm{t}_1\in\mathrm{T}$, un vecteur de covariables x, on veut estimer la CATE [13] :

$$\tau(\mathrm{t}_0,\mathrm{t}_1,\mathrm{x}) = \mathrm{E}[Y(\mathrm{t}_1)-Y(\mathrm{t}_0)|X=\mathrm{x}] \qquad (\mathrm{CATE})$$

On peut réécrire les données sous la forme : $\{Yi(Ti), Ti, Xi, Wi, Zi\}$, où $Yi(Ti)$ est l'outcome observé pour un traitement donné, Ti est le traitement, Xi sont les covariables utilisées pour le conditionnement de CATE, Wi sont les autres covariables qui affectent quand l'outcome $Yi(Ti)$ et potentiellement le traitement Ti et Zi sont les variables qui affectent le traitement Ti mais n'affectent pas directement l'outcome $Yi(Ti)$ [13].

On peut voir nos données comme des échantillons indépendants et identiquement distribués $\{Yi(Ti), Ti, Xi, Wi, Zi\}$ suivant le modélisé par le système suivant :

$$Y = g(T, X, W, \epsilon)$$

$$T = f(X, W, Z, \eta)$$

Où ϵ et η sont des bruits indépendants de X, Z, T, W. On peut alors réécrire la CATE comme [13] :

$$\tau(t_0, t_1, x)=E[g(t_1, X, W, \epsilon)-g(t_0, X, W, \epsilon)|X=x] \quad (CATE)$$

Choix du modèle utilisé

Puisqu'on se situe dans le cas d'une étude non-randomisée (autrement dit, on a pas attribué aléatoirement les clefs conso-live à nos clients), et qu'on ne dispose a priori pas de variable instrumentale qui affecterait le choix du traitement mais pas l'outcome, il est nécessaire d'introduire un maximum de covariables/variables de contrôle permettant de décrire le choix du traitement afin de réduire le biais au maximum.

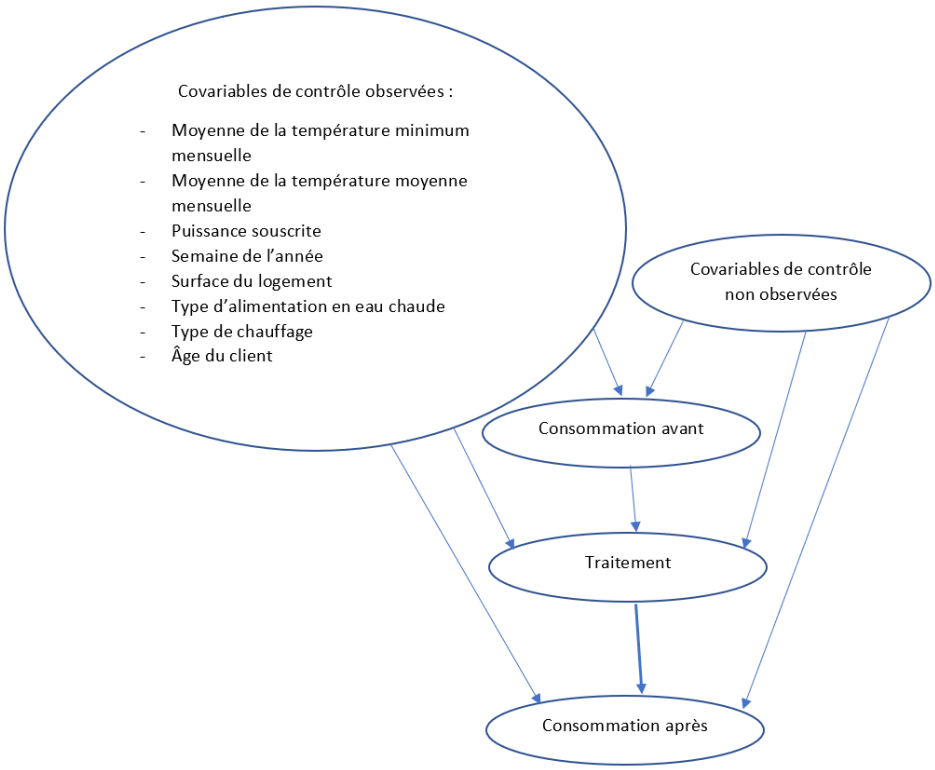


Figure 18- Schéma de relations entre les différentes variables

Malgré cet enrichissement du modèle, on reste dans le cas où on a assez peu de variables influençant la réponse au traitement. Aussi, on peut se permettre de ne pas utiliser les modèles Sparses.

Enfin, une hypothèse assez courante, suggérée comme première approche par la librairie EconLM afin d'utiliser les modèles les plus simples d'accès et gratuite dans le cas binaire [13], est la linéarité de l'hétérogénéité par rapport au traitement. C'est-à-dire :

$$Y = H(X, W) \cdot T + g(X, W, \epsilon)$$

$$T = f(X, W, Z, \eta)$$

$H(X, W)$ est alors la CATE (aussi notée $\theta(X)$).

La Flow-Chart[Figure 25] de sélection de modèle proposée par EconML indique alors que dans notre cas, il faut utiliser le modèle LineardDML ou LinearDRLearner.

- LinearDML [14] : Linear Double Machine Learning est un algorithme qui repose sur une double régression : la première afin d'estimer $Y' = Y - \mathrm{E}[Y|X, W]$ et la seconde afin d'estimer $T' = T - \mathrm{E}[T|X, W]$

  On a alors : $\theta' = \mathrm{argmin}_{\theta \in \Theta} \mathrm{E}[(\mathrm{Y}' - \theta(\mathrm{X}) \cdot \mathrm{T}')^2]$
  Les méthodes de régressions sur Y' et T' sont des degrés de liberté laissés à l'utilisateur.

- LinearDRLearner [15] : Linear Doubly Robust Learner est un algorithme reposant lui aussi sur la double régression de Y' et T' cependant, ici, le modèle prédit Y' comme une fonction de T. Un tel modèle sert notamment à trouver la politique de traitement optimale. Aussi, nous ne sommes pas concernés par ce cas et nous nous intéresserons plus à LinearDML.

Analyse d'impact

Afin de pouvoir comparer l'impact de conso-Live sur la consommation des clients, on cherche donc à construire un dataset de la manière suivante :

- On prend des clients ayant la clef conso-live depuis plus d'un an et on enregistre leurs informations l'année précédant l'utilisation de conso-live et l'année suivant l'utilisation de conso-live
- Pour chaque client A dans le groupe précédent, on trouve un client B témoin qui ne dispose pas de la clef conso-Live dont on enregistre les informations aux mêmes dates que le client A afin d'avoir deux groupes de tailles semblables.

Malheureusement, la première étape de ce procédé nécessite de poser une contrainte d'un an à gauche et d'un an à droite sur l'historique des clients. Aussi, on perd beaucoup de clients potentiels dans ce filtre (20 000 clients restants).

Après avoir créé un modèle LinearDML composé de deux régressions de Lasso pour le calcul de Y' et T', on obtient les résultats suivants :

Uncertainty of Mean Point Estimate :

- Mean point : -0.441
- Pvalue : 0.25
- CI_lower_mean : -1.072
- CI_upper_mean : 0.189

Autrement dit, la clef conso-Live réduirait en moyenne de 0.441 kWh la consommation d'électricité l'année suivant l'acquisition de la clef.

Ce résultat est assez décevant : outre la réduction assez légère, on obtient une p-value très haute et donc une conclusion à laquelle on ne peut pas se fier du tout (l'intervalle de confiance monte jusqu'à -1.072, ce qui signifierait que potentiellement la clef conso-Live augmenterait la consommation des clients).

Ce mauvais résultat peut être attribué à plusieurs facteurs :

- Un manque de clients avec suffisamment d'historique
- Un manque de co-variables de contrôle, il faudrait idéalement enrichir le modèle de plusieurs autres variables impactant le traitement et/ou la consommation

- L'utilisation encore expérimentale de cette bibliothèque d'économétrie. Avec plus de temps, il aurait probablement été possible de mieux tuner le modèle ou de sélectionner des sous-modèles plus appropriés
- La manière dont le problème a été posé. On pourrait notamment redéfinir le traitement et l'output de manière à ne s'intéresser qu'au mois suivant/précédent l'utilisation de conso-live. On résoudrait ainsi le problème de gros historique nécessaire d'un client pour le faire participer à l'étude.

Prédiction des étiquettes énergétiques

Le Diagnostic de Performance Energétique [16] (DPE) renseigne sur la performance énergétique d'un logement ou d'un bâtiment, en évaluant sa consommation d'énergie et son impact en termes d'émissions de gaz à effet de serre. Il s'inscrit dans le cadre de la politique énergétique définie au niveau européen afin de réduire la consommation d'énergie des bâtiments et de limiter les émissions de gaz à effet de serre.

C'est une information importante qui pourrait être utilisée par de nombreux algorithmes de machine learning développés par TotalEnergies notamment pour proposer à des clients thermosensibles des travaux financés par l'Etat dans le but d'améliorer leur étiquette.

Toutefois, si on dispose de l'étiquette énergétique de certains clients, ce n'est pas le cas en général. Il est donc intéressant de réussir à prédire cette étiquette de manière automatique en fonction des autres informations dont nous disposons sur nos clients.

On se concentre ici sur les clients chauffés à l'électricité, la majorité du parc, qui sont suffisamment nombreux pour pouvoir faire tourner des algorithmes de machine learning.

Approche empirique

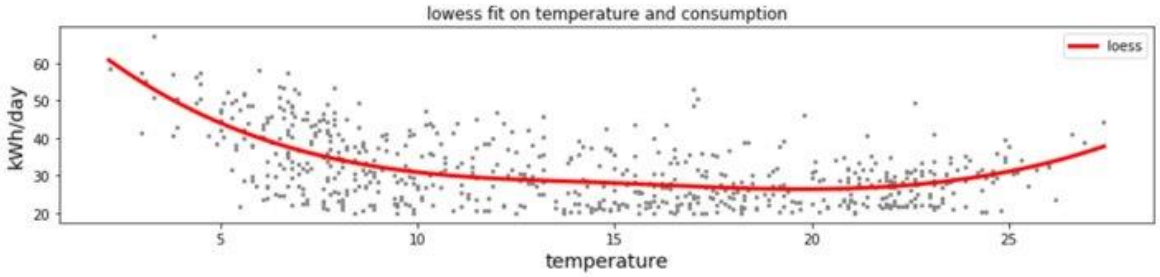


Figure 19 - Consommation d'un client en fonction de la température

Naturellement, un client a un logement fortement thermosensible lorsque :

- Il chauffe plus tôt que des clients avec un profil similaire au sien (il commence à chauffer à 14°C par exemple)
- Il chauffe plus que la moyenne des clients similaires pour compenser chaque degré extérieur perdu (pente plus raide à gauche)

Cependant, on peut attribuer ce résultat à plusieurs facteurs :

- Mauvaise performance énergétique du logement
- Appareils supplémentaires (piscine chauffée, sauna, …) qui n'est pas pris en compte dans lors de la création des groupes de clients similaires
- Comportement différent (degré de confort différent, mauvaise utilisation du chauffage, …)

Par conséquent, on ne peut pas se contenter d'étudier la forme du nuage de points de la consommation en fonction de la température, comme on l'avait fait pour le bilan hiver.

Approche Statistique

On fait la jointure avec les informations dont TotalEnergies dispose sur le logement (ancienneté du logement, département, type de chauffage, type de chauffe-eau, travaux de rénovation énergétique, …), la consommation (pente et température de début de chauffe mentionnées plus haut, corrélation de pearson et de spearman du nuage de points, …) et le contrat du client (option tarifaire, puissance souscrite, type de contrat : location/propriété …).

Ces informations sont fortement corrélées[Figure 26 - Corrélation des features avec les étiquettes énergétiques] à la qualité de l'étiquette mais les différences d'une étiquette à l'autre sont trop fines pour pouvoir prédire directement les étiquettes. On préfère se fier à un modèle plus simple qui va trier les clients en deux groupes : bonne étiquette énergétique (ABC : 0) et mauvaise étiquette énergétique (DEFG : 1).

Après avoir resamplé les données de manière à avoir autant de bonnes étiquettes que de mauvaises, on crée un modèle de classification binaire supervisée (ici on utilise encore une fois LightGBM [17], une version de XGBoost allégée en calculs).

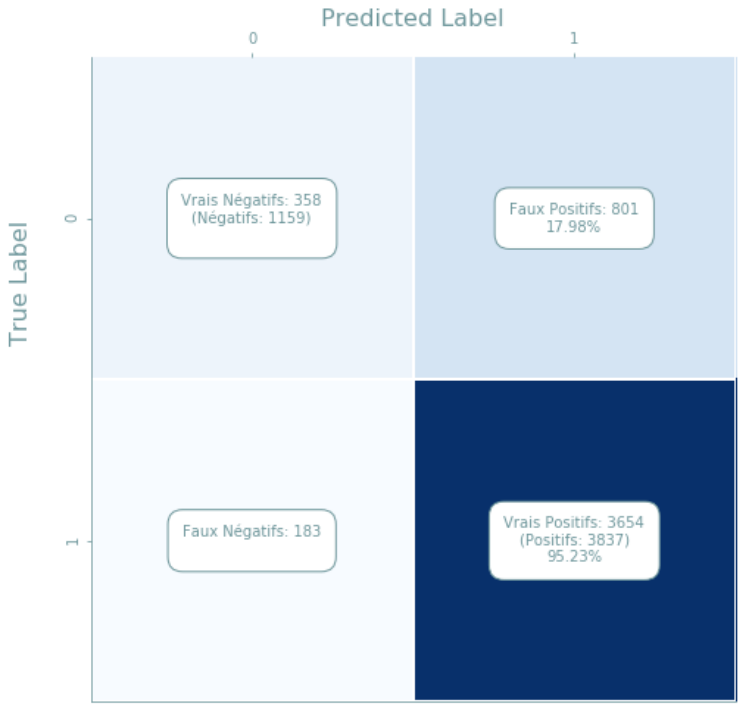


Figure 20 - Matrice de confusion étiquettes DPE

En resamplant les données, on a augmenté la concentration de mauvaises étiquettes dans le jeu de train, aussi le modèle a particulièrement bien réussi à les reconnaitre (Rappel de 95.23%) mais on a aussi introduit au passage plus de Faux Positifs (18% des étiquettes ABC ont été considérées comme de mauvaises étiquettes par notre modèle). Dans notre cas, ce n'est pas un problème si important dans la mesure où un client faux positif recevra un mail lui proposant des travaux de rénovations qui ne l'intéressera pas forcément.

En revanche, il est intéressant d'étudier de plus près les variables qui ont pu induire notre modèle en erreur. Pour ce faire, on utilise les shap values (concept emprunté à la théorie du jeu qui permet de calculer l'impact de chaque variable sur l'output d'un modèle) et notamment la librairie python shap dédiée à la compréhension des modèles de machine learning afin de diminuer l'effet « blackbox » de nos résultats [18]:

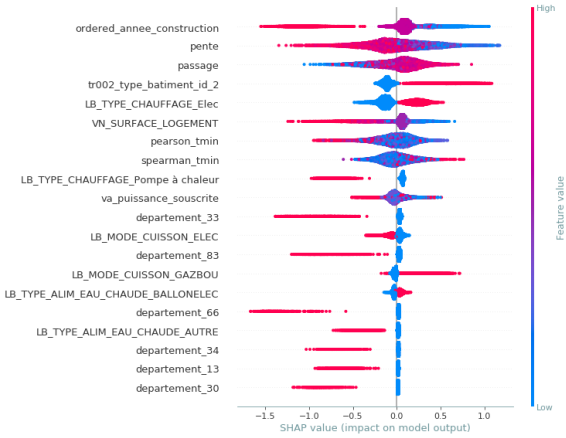


Figure 21 - Shap Values de la population

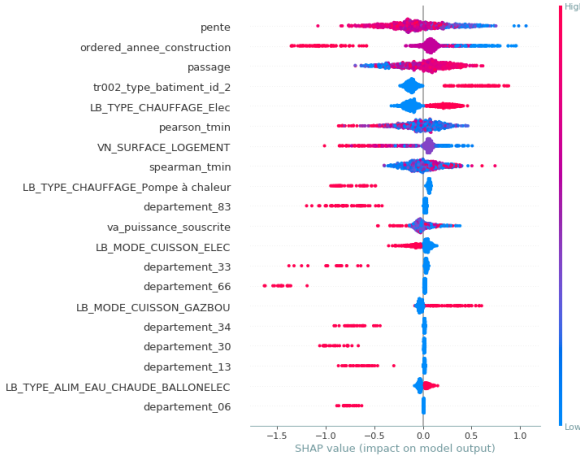


Figure 21 - Shap Values des faux positifs

On lit le graphique de la manière suivante :

- Les features sont ordonnées verticalement par importance d'impact sur l'output (la pente du nuage de points et l'année de construction sont les features les plus importantes)
- Pour chaque feature, un nuage de points est affiché, représentant l'impact que la feature a eu sur l'output par son placement horizontal (un point très à gauche signifie que pour un client, la feature a eu un impact très fort sur le fait que le client soit considéré comme ayant une bonne étiquette : 0 et inversement, un point très à droite signifie que la feature est un indice que le client est thermosensible : 1)
- La couleur représente la valeur de la feature pour chaque client (attention, dans le cas des pentes, les valeurs sont négatives, aussi les fortes pentes sont en bleu)

Malheureusement, ici, on ne voit pas se dégager de tendance évidente entre la population étudiée et l'ensemble des faux-positifs. La librairie shap permet cependant de faire du peaking et de regarder au cas par cas quels ont été les features les plus importantes dans le choix de l'output :

Figure 24 - exemple de force_plot d'un client faux positif

En regardant de plus près, client par client, on voit que le modèle a tendance à considérer les logements des clients ayant une bonne étiquette DPE comme énergivore quand :

- La pente de leur nuage de points est trop forte
- Ils commencent à chauffer plus tôt que les autres clients (voir la variable « passage »)

- La taille de leur logement est assez importante
- Quand ils ne possèdent pas de pompe à chaleur

Dans l'exemple plus haut, le modèle estime à 95% la probabilité que le client soit dans un logement thermosensible, alors qu'il a probablement des habitudes de vie énergivores.

L'analyse de ce genre de client permet de soulever une nouvelle piste de travail : contacter les clients que notre modèle considère comme thermosensibles alors qu'ils ont une bonne étiquette DPE à cause de leur consommation afin de les sensibiliser à des modes de vie moins énergivores.

Les résultats de cette prédiction, ainsi que les caractéristiques des clients classés comme faux-positifs, ont été présentés au service Marketing qui a été fortement intéressé par la possibilité d'utiliser ces résultats afin d'envoyer des propositions de travaux ou des conseils de maitrise d'énergie à nos clients.

Conclusion

Ce stage a été pour moi l'occasion de travailler sur plusieurs projets en parallèle, pour certains en interaction directe avec d'autres professionnels de TotalEnergies. J'ai beaucoup appris sur les méthodes de travail dans le milieu de l'entreprise et j'ai notamment pu me familiariser avec un fonctionnement par sprint via l'utilisation de la méthode agile.

J'ai apprécié le fait de travailler autour de sujets ayant du sens (particulièrement ceux qui visaient à accompagner le client vers une meilleure maitrise de son énergie) tout en développant de nouvelles compétences en DataScience en particulier avec des outils comme PySpark et Databricks qui sont des standards pour la production de code en entreprise.

Je suis particulièrement fier d'avoir pu terminer de nombreux projets débutés au cours de ce stage :

- Prédiction d'appels – Pro
- Prédiction d'appels – Ventes (pour lequel j'ai quasiment travaillé en autonomie)
- Bilan hiver (gaz)
- Prédiction d'étiquettes DPE
- Traduction de la loess en PySpark

Ce stage m'a permis de découvrir le métier de DataScientist et de me conforter dans l'idée que c'était la voie que je comptais emprunter à la fin de mes études.

Remerciements

Si ce stage s'est aussi bien passé, c'est grâce à l'accueil et à l'encadrement que j'ai reçu. Aussi, je tenais à remercier personnellement toute l'équipe Data de son accueil et de sa prévenance.

Je tiens à remercier chaleureusement toutes les personnes qui ont contribué au succès de mon stage.

Je voudrais dans un premier temps remercier ma directrice de stage, Eurydice Lafferayrie, manager de l'équipe Data au début de mon stage, pour m'avoir offert cette belle opportunité, pour son accueil, son encadrement hebdomadaire, son écoute, sa disponibilité et ses conseils.

Ensuite, je voudrais remercier mon encadrant Yoan Saint-Pierre, DataScientist de TotalEnergies, pour la manière dont il a axé tous les projets qu'il m'a confié, selon leur intérêt, dans le cadre de mon stage et de mon profil de futur DataScientist, pour la confiance qu'il m'a manifesté en m'accordant une large indépendance sur le projet de la prévision d'appels pour les ventes, également pour la pédagogie dont il a su faire preuve tout au long du stage, pour sa patience, ses nombreux conseils et son accueil très chaleureux.

Je remercie également Thomas Jassem, DataScientist du pôle BigData, pour son accueil chaleureux et pour ses nombreux conseils au cours des réunions hebdomadaires de partage des avancées de projets de DataScience. En particulier, je le remercie de m'avoir partagé son expertise de PySpark.

Je remercie aussi Sophie Charvet, cheffe du projet AVA (conso-live) au pôle Marketing pour ses explications détaillées de l'entreprise à mon arrivée, sa pédagogie, son accueil très humain et ses efforts vis-à-vis de mon intégration inter-services.

Enfin, je remercie chacun des membres de l'équipe du pôle Data de TotalEnergies pour leur accueil et pour le contact humain très présent qu'ils ont maintenu, et ce malgré la période de confinement et le travail en distanciel. J'ai beaucoup appris sur le travail en entreprise à leurs côtés.

Bibliographie

[1] «EkWateur - histoire du marché de l'énergie,» [En ligne]. Available: https://ekwateur.fr/2020/05/15/histoire-marche-energie/.

[2] «Ecologie.gouv,» [En ligne]. Available: https://www.ecologie.gouv.fr/tarifs-gaz.

[3] «TDE - histoire du fournisseur,» [En ligne]. Available: https://opera-energie.com/fournisseur-energie-direct-energie/.

[4] «totalenergies.com part de marché,» [En ligne]. Available: https://totalenergies.com/fr/medias/actualite/communiques-presse/total-direct-energie-depasse-5-millions-clients-france.

[5] «spark.apache.org,» [En ligne]. Available: https://spark.apache.org/.

[6] «Pyspark Documentation,» [En ligne]. Available: http://spark.apache.org/docs/latest/api/python/.

[7] «Databricks - homepage,» [En ligne]. Available: https://databricks.com/fr/.

[8] «Medium - Math of Prophet,» [En ligne]. Available: https://medium.com/future-vision/the-math-of-prophet-46864fa9c55a.

[9] «Facebook Prophet - homepage,» [En ligne]. Available: https://facebook.github.io/prophet/.

[10] «Neural Prophet - github,» [En ligne]. Available: https://github.com/ourownstory/neural_prophet.

[11] «Wikipedia - Loess,» [En ligne]. Available: https://en.wikipedia.org/wiki/Local_regression.

[12] G. Pauline, «Méthodes économétriques pour l'évaluation de politiques publiques,» 2014. [En ligne]. Available: https://www.cairn.info/revue-economie-et-prevision-2014-1-page-1.htm .

[13] «EconML - Documentation,» [En ligne]. Available: https://econml.azurewebsites.net/spec/api.html.

[14] «EconML - LinearDML documentation,» [En ligne]. Available: https://econml.azurewebsites.net/spec/estimation/dml.html.

[15] «EconML - LinearDRLearner documentation,» [En ligne]. Available: https://econml.azurewebsites.net/spec/estimation/dr.html.

[16] «Ecologie.gouv - DPE,» [En ligne]. Available: https://www.ecologie.gouv.fr/diagnostic-performance-energetique-dpe.

[17] «LightGBM - homepage,» [En ligne]. Available: https://lightgbm.readthedocs.io/en/latest/.

[18] «Shap - documentation,» [En ligne]. Available: https://shap.readthedocs.io/en/latest/index.html.

Annexe

Analyse d'impact

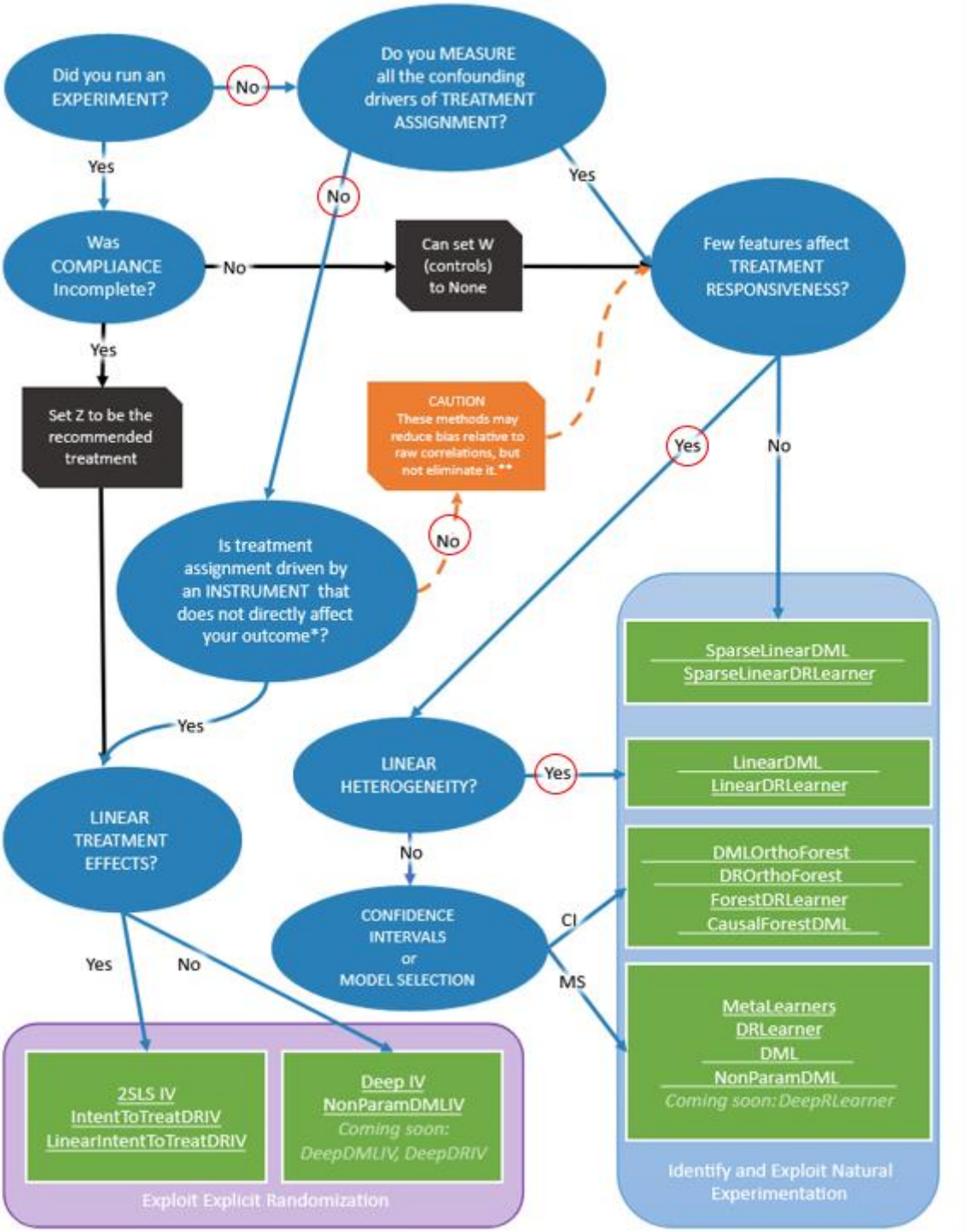
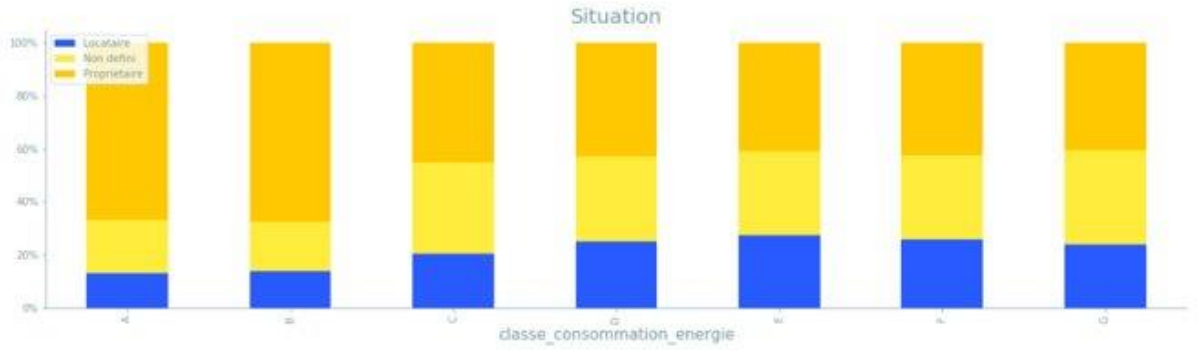
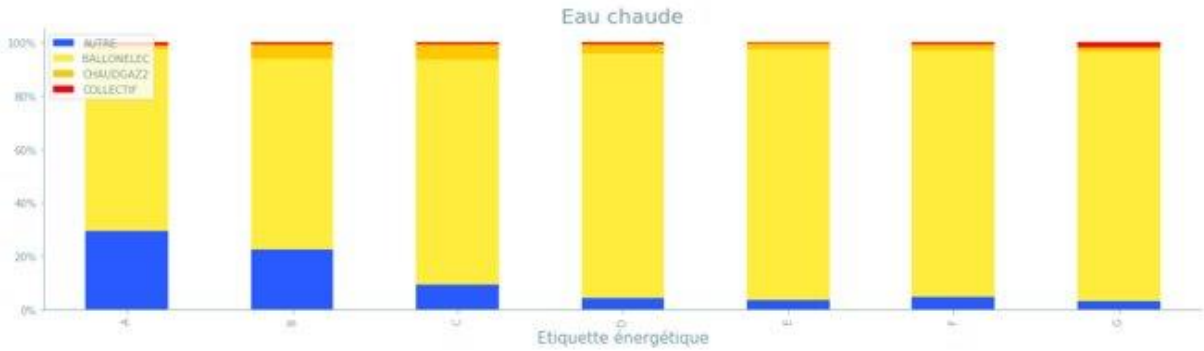
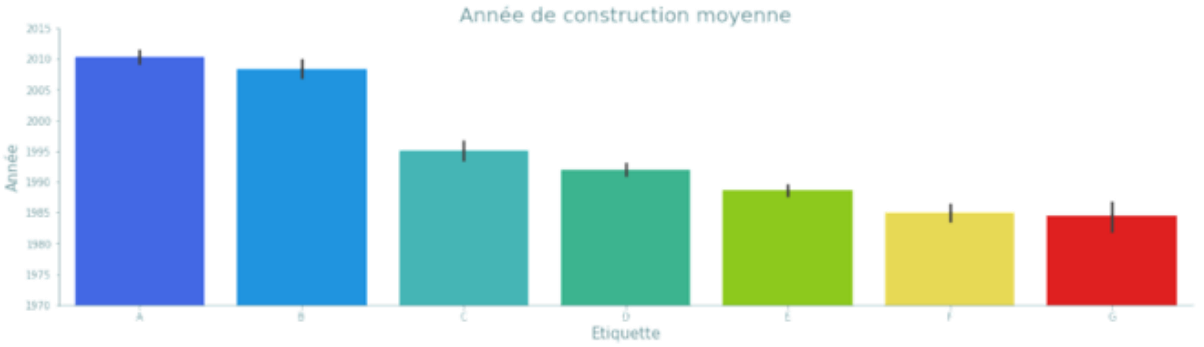


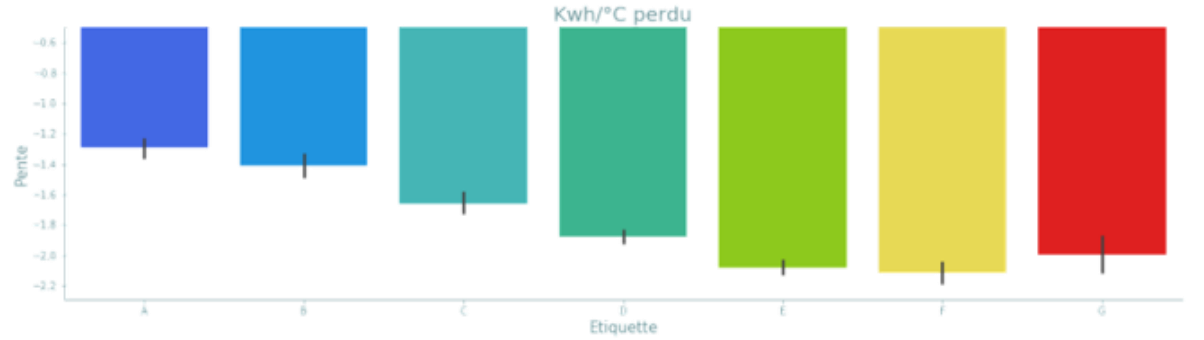
Figure 25- Flow Chart choix du modèle EconML

Etiquette énergétique









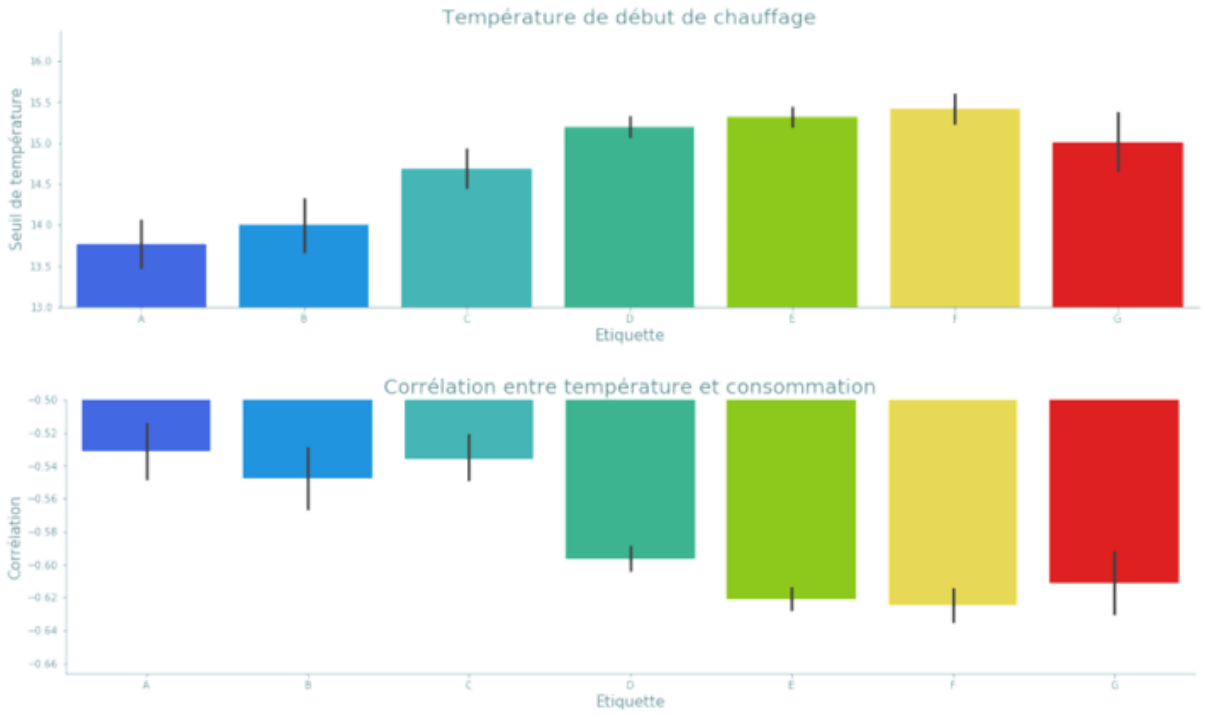


Figure 26 - Corrélation des features avec les étiquettes énergétiques