Machine Learning pour expliquer le prix de l'électricité

1 Contexte

Une multitude de facteurs influencent le prix de l'electricité au quotidien. Des variations locales du climat pourront à la fois affecter la production et la demande électrique par exemple. Des phénomènes à plus long terme, comme le réchauffement climatique, auront également un impact évident. Des évènements géopolitiques, comme la guerre en Ukraine, peuvent en parallèle faire bouger le prix des matières premières qui sont clefs dans la production d'électricité, sachant que chaque pays s'appuie sur un mix énergétique qui lui est propre (nucléaire, solaire, hydrolique, gaz, charbon, etc). De plus chaque pays peut importer/exporter de l'électricité avec ses voisins au travers de marchés dynamiques, comme en Europe. Ces différents élements rendent assez complexe la modélisation du prix de l'électricité par pays.

2 Objectif du projet

L'objectif est de modéliser le prix l'électricité à partir de données météorologiques, énergétiques (matières premières) et commerciales pour deux pays européens - la France et l'Allemagne.
Il est à noter qu'il s'agit plutôt d'un problème d'explication des prix par d'autres variables simultanées et non pas d'un simple problème de prédiction. Plus précisément le but est de construire un modèle qui, à partir de ces variables explicatives, renvoie une bonne estimation de la variation journalière du prix de contrats à terme (dits **futures**) sur l'électricité, en France ou en Allemagne. Ces contrats permettent d'acheter (ou de vendre) une quantité donnée d'électricité à un prix fixé par le contrat et qui sera livrée à une date future spécifiée (maturité du contrat). Les **futures** sont donc des instruments financiers qui donnent une estimation de la valeur de l'électricité au moment de la maturité du contrat à partir des conditions actuelles du marché. Dans le cadre de ce projet, on se restreint à des **futures** à courte maturité (24h). Il est important de noter que l'échange de futures sur l'electricité est un marché dynamique en Europe.
Concernant les variables explicatives, les participants auront accès pour chaque pays à des mesures journalières de données météorologiques (température, quantité de pluie et force du vent), de production énergétique (variations des prix de différentes matières premières / énergies) et d'utilisation de l'électricité (consommation, échanges entre ces deux pays, import-export avec le reste de l'Europe).

3 Etapes du projet

La méthode CRISP [1] (initialement connue comme CRISP-DM) a été au départ développée par IBM dans les années 60 pour réaliser les projets Data-mining. En Data Science, elle demeure la méthodologie la plus utilisée. Elle se décompose en 6 étapes allant de la compréhension du problème métier au déploiement et la mise en production. Cette méthode est agile et itérative, c'est-à-dire que chaque itération apporte de la connaissance

1. CRISP-DM : Cross-Industry Standard Process for Data Mining. Cette annexe est issue de https ://www.mygreatlearning.com/blog/why-using-crisp-dm-will-make-you-a-better-data-scientist

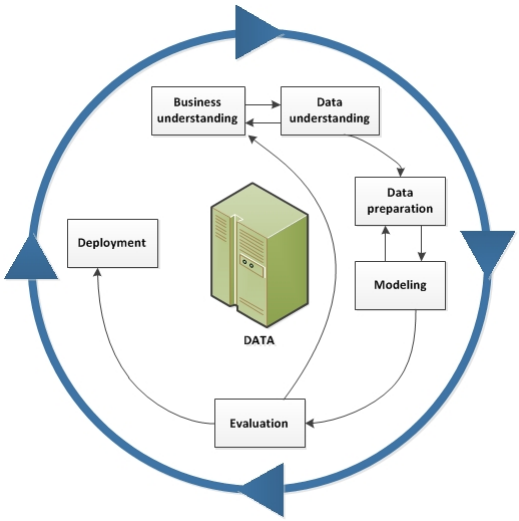

FIGURE 1 – Les phases de CRISP-DM

métier supplémentaire qui permet de mieux aborder l'itération suivante. Nous adoptons dans ce qui suit cette méthodologie. Ainsi, n'oubliez pas de réfléchir à l'issue de chaque itération et mettre à jour si nécessaires l'étape précédente.

3.1 Description des données

Trois jeux de données sont fournis au format csv :

- les données en entrée **Data_X** et en sortie **Data_Y**,
- Nouvelles données en entrée non labélisées : **DataNew_X**.

Les données en entrée **Data_X** et **DataNew_X** représentent les mêmes variables explicatives mais sur deux périodes différentes.
Les données d'entrée possèdent 35 colonnes :

- ID : Identifiant d'indexe unique, associé à un jour (DAY_ID) et un pays (COUNTRY),
- DAY_ID : Identifiant du jour - les dates ont été annonymisées en préservant la structure des données,
- COUNTRY : Identifiant du pays - DE = Allemagne, FR = France,

et composées ensuite de variations journalières du prix de matières premières,

- GAS_RET : Gaz en Europe,
- COAL_RET : Charbon en Europe,
- CARBON_RET : Futures sur les emissions carbone,

de mesures météorologiques (journalières, dans le pays x),

- x_TEMP : Temperature,
- x_RAIN : Pluie,
- x_WIND : Vent,

de mesures de productions d'energie (journalière, dans le pays x),

- x_GAS : Gaz naturel,
- x_COAL : Charbon,
- x_HYDRO : Hydrolique,
- x_NUCLEAR : Nucléaire,
- x_SOLAR : Photovoltaïque,
- x_WINDPOW : Eolienne,
- x_LIGNITE : Lignite,

et de mesures d'utilisation électrique (journalières, dans le pays x),

- x_CONSUMPTON : Electricité totale consommée,
- x_RESIDUAL_LOAD : Electricité consommée après utilisation des énergies renouvelables,

— x_NET_IMPORT : Electricité importée depuis l'Europe,
— x_NET_EXPORT : Electricité exportée vers l'Europe,
— DE_FR_EXCHANGE : Electricité échangée entre Allemagne et France,
— FR_DE_EXCHANGE : Electricité échangée entre France et Allemagne.

Les données en sortie se composent de deux colonnes :

— ID : Identifiant unique - le même que celui des données d'entrée,
— TARGET : Variation journalière du prix de futures d'électricité (maturité 24h).

3.2 Préparation des données

En fonction du type du problème et les objectifs à atteindre, la préparation des données comporte généralement les tâches suivantes :

— Fusion des ensembles et/ou enregistrements de données
— Sélection d'un sous-ensemble de données
— calcul de nouveaux attributs
— Tri des données en vue de la modélisation
— Suppression ou remplacement des blancs ou des valeurs manquantes
— Fractionnement en sous-ensembles d'apprentissage et de test

Dans le cadre de ce projet, nous vous demandons en particulier de :

— Vérifier s'il y a des valeurs manquantes dans les données.
— Vérifier si les valeurs des différents attributs sont comparables

Remarque. Vous pouvez toujours proposer d'autres traitements mais il faudra expliquer à chaque fois. Pour bien maîtriser cette phase, je vous invite à consulter le lien suivant : Data Preparation with pandas

3.3 Analyse exploratoire des données

Cette phase consiste à :

1. identifier la variable cible qui doit être prédite. Dans ce projet la variable prédite est la variation journalière des prix des futures (la colonne TARGET dans les datasets Y_train et Y_test).
2. effectuer une analyse exploratoire des données (EDA, Exploratory Data Analysis) par une variété de graphiques et de statistiques en suivant les étapes suivantes :
   — Faire un aperçu des variables en examinant leur type, leur distribution, leur plage de valeurs et leur signification
   — Examiner la relation entre les variables caractéristiques et la variable cible en utilisant des techniques graphiques telles que des histogrammes, des diagrammes en boîte et des graphiques de dispersion
   — Construire une matrice de corrélation entre les variables
   — Interpréter les résultats de l'EDA pour identifier les caractéristiques importantes qui influencent le prix de l'électricité et les relations significatives entre les variables

Pour bien maîtriser cette phase, je vous invite à consulter le lien suivant : Python for Data Science: Implementing Exploratory Data Analysis (EDA) and K-Means Clustering

3.4 Modélisation des données

Différents algorithmes de Machine Learning seront utilisés pour entraîner des modèles de prédiction à partir des données. Nous proposons d'implémenter les six modèles de régression :

1. Régression linéaire simple
2. Régression linéaire régularisée (régression RIDGE / régression LASSO) : voir lien suivant Régression Ridge, Lasso et nouvel estimateur
3. Méthode des k plus proches voisins pour la régression (K-NN, k-Nearest Neighbors regressor)
4. Arbres de décision pour la régression (Decision tree regressor)
5. **En bonus**, les Forêts aléatoires (Random Forest regressor)

Vous devez comprendre les deux variantes de régression linéaire régularisée et les décrire brièvement dans votre rapport. Pareil pour la méthode bonus.

3.5 Evaluation des modèles

Les modèles seront évalués en utilisant des métriques relatives à la régression telles que :

- la corrélation de Spearman,
- le coefficient de détermination R2
- l'erreur quadratique moyenne (RMSE).

Au cours de cette phase vous devez :

1. Optimiser chaque modèle : Varier les hyperparamètres de chaque modèle et retenir ceux qui aboutissent aux meilleurs performances.
2. Comparaison des différents modèles : Comparer les performances des algorithmes de prédiction des données et choisir le plus performant en effectuant un classement.
3. Pour le meilleur modèle retenu, évaluer l'importance des variables (attributs) qui ont about à la meilleure prédiction (Evaluez l'importance des variables, How to Calculate Feature Importance With Python, Feature importance, Feature Importance Explained)

4 Langage de programmation et choix des bibliothèques

- **Langage de programmation :** Python
- **Choix des bibliothèques**
  - **Manipulation des données :** Pandas
  - **Visualisation des données :** Matplotlib & Seaborn
  - **Modélisation des données :** Scikit-learn

5 Livrables

- Des notebooks Jupyter (vous pouvez utiliser Goggle Colab Notebook - Google Colaboratory Notebooks) permettant de voir et suivre tout le travail (il doit comporter une explication des méthodes ML non vues en cours, les résultats expérimentaux et leurs interprétations).
- **Ou bien** une archive des scripts pythons et un rapport succinct qui décrit votre projet (il doit comporter une explication des méthodes ML non vues en cours, les résultats expérimentaux et leurs interprétations).
- une présentation pptx de 10 minutes

6 Echéances

1. Dépôt du projet sur moodle : 7 avril 2022
2. Durée du projet : un mois
3. Présentation du projet en cours et réponses à des questions : Semaine du 11 avril 2023
4. Suivi intermédiaire : Semaine 16
5. Date limite de remise des travaux sur moodle : 8 mai 2023 à 23h55. Une seule copie par equipe.
6. Soutenance : Au cours de la séance de la semaine 19.

7 Barème

Votre projet sera noté par votre prof pendant la dernière séance du cours (séance de la semaine 19) en se basant sur votre présentation et le travail fournit.

1. Rapport (clarté, interprétations) : 30%
2. Code (répond aux besoins et qualité) : 35%
3. Soutenance et réponse aux questions : 35%

Remarque. Les membres d'une même équipe peuvent ne pas avoir la même note et cela en fonction de leur degré d'investissement dans le projet. Une petite note contenant la description du degré d'investissement de chaque membre de l'équipe sous forme de pourcentage doit être communiquée à l'enseignant le jour de la soutenance.

8 Accès à la plateforme nowledgeable

Nous mettrons à votre disposition un ensemble d'exercices via la plateforme nowledgeable afin de maîtriser les différentes étapes d'un projet datascience : Efrei - Machine learning L3 - projet