

Anticipez les besoins en consommation électrique de bâtiments

MISSION

Vous travaillez pour la **ville de Seattle**. Pour atteindre son objectif de ville neutre en émissions de carbone en 2050, votre équipe s'intéresse de près aux émissions des bâtiments non destinés à l'habitation.



Seattle

Les données

Les données de consommation sont à télécharger à [cette adresse](#).

Problématique de la ville de Seattle

Des relevés minutieux ont été effectués par vos agents en 2015 et en 2016. Cependant, ces relevés sont coûteux à obtenir, et à partir de ceux déjà réalisés, **vous voulez tenter de prédire les émissions de CO2 et la consommation totale d'énergie** de bâtiments pour lesquels elles n'ont pas encore été mesurées.

Votre prédiction se basera sur les données déclaratives du permis d'exploitation commerciale (taille et usage des bâtiments, mention de travaux récents, date de construction..)

Vous cherchez également à **évaluer l'intérêt de l'"[ENERGY STAR Score](#)" pour la prédiction d'émissions**, qui est fastidieux à calculer avec l'approche utilisée actuellement par votre équipe.

Votre mission

Vous sortez tout juste d'une réunion de brief avec votre équipe. Voici un récapitulatif de votre mission :

1. Réaliser une courte analyse exploratoire.
2. Tester différents modèles de prédiction afin de répondre au mieux à la problématique.

Avant de quitter la salle de brief, **Douglas**, le **project lead**, vous donne quelques pistes, et erreurs à éviter :

L'objectif est de te passer des relevés de consommation annuels (attention à la fuite de données), mais rien ne t'interdit d'en déduire des variables plus simples (nature et proportions des sources d'énergie utilisées).

Fais bien attention au traitement des différentes variables, à la fois pour trouver de nouvelles informations (peut-on déduire des choses intéressantes d'une simple adresse ?) et optimiser les performances en appliquant des transformations simples aux variables (normalisation, passage au log, etc.).

Mets en place une évaluation rigoureuse des performances de la régression, et optimise les hyperparamètres et le choix d'algorithme de ML à l'aide d'une validation croisée.

Livrables attendus

- Un **notebook** de l'analyse exploratoire mis au propre et annoté.
- Le **code** (ou un notebook) des différents tests de modèles mis au propre, dans lequel vous identifierez clairement le modèle final choisi.
- Un support de **présentation** pour la soutenance.

Pour faciliter votre passage au jury, déposez sur la plateforme, dans un dossier nommé "*Pélec_nom_prenom*", tous les livrables du projet. Chaque livrable doit être nommé avec le numéro du projet et selon l'ordre dans lequel il apparaît, par exemple "*Pélec_01_notebook*", "*Pélec_02_code*", et ainsi de suite.

Modalités de la soutenance

5 min - Présentation de la problématique, de son interprétation et des pistes de recherche envisagées.

5 min - Présentation du cleaning effectué, du feature engineering et de l'exploration.

10 min - Présentation des différentes pistes de modélisation effectuées.

5 min - Présentation du modèle final sélectionné ainsi que des améliorations effectuées.

5 à 10 minutes de questions-réponses.

Compétences évaluées

- Mettre en place le modèle d'apprentissage supervisé adapté au problème métier
- Adapter les hyperparamètres d'un algorithme d'apprentissage supervisé afin de l'améliorer
- Transformer les variables pertinentes d'un modèle d'apprentissage supervisé
- Évaluer les performances d'un modèle d'apprentissage supervisé