Prédiction de l'empreinte énergétique de Seattle

Un projet data-driven pour accompagner la ville vers la neutralité carbone à l'horizon 2050

VISION STRATÉGIQUE

L'ambition 2050 : neutralité carbone

Seattle s'est fixée un objectif ambitieux : atteindre la neutralité carbone d'ici 2050. Pour y parvenir, la ville doit transformer radicalement son parc immobilier non-résidentiel, responsable d'une part significative des émissions urbaines.

Ce projet propose une approche proactive innovante : prédire la consommation énergétique dès l'obtention du permis d’exploitation commerciale. Cette anticipation permettra d'optimiser les politiques d'efficacité énergétique avant même la mise en service des bâtiments.
L'enjeu est considérable : identifier les leviers d'action les plus pertinents pour accompagner les acteurs de la construction vers des choix énergétiques optimaux.

2050

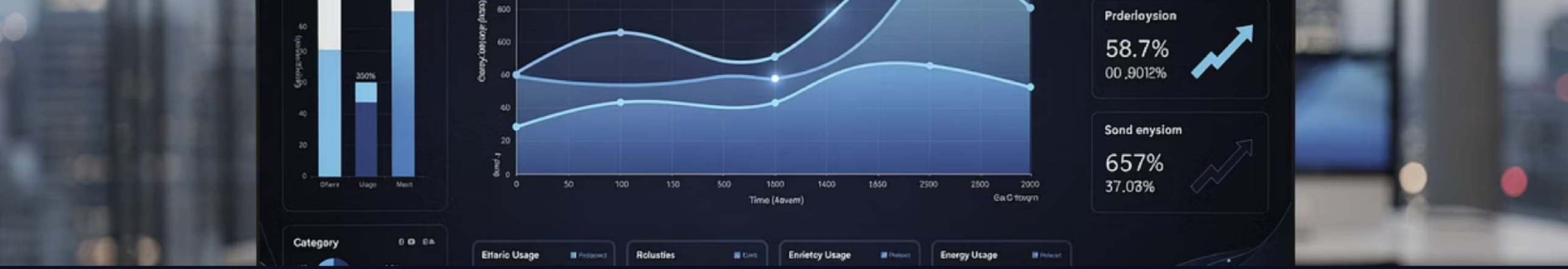
Objectif neutralité

Échéance fixée par la ville

100%

Parc non-résidentiel

Périmètre du projet



La question centrale de recherche

> Peut-on prédire avec précision la consommation énergétique d'un bâtiment uniquement à partir de ses caractéristiques physiques ? Et quel est l'impact réel du score ENERGY STAR dans cette équation ?

Cette double interrogation structure l'ensemble de notre démarche. D'une part, nous cherchons à quantifier la capacité prédictive des attributs structurels (surface, usage, âge). D'autre part, nous évaluons la valeur ajoutée du score ENERGY STAR, cet indicateur standardisé de performance énergétique.

La réponse à ces questions permettra aux décideurs municipaux d'identifier les variables clés à collecter et les certifications à promouvoir pour maximiser l'efficacité des politiques publiques.

DONNÉES

Préparation des données

01

Ciblage du périmètre

Exclusion délibérée du secteur résidentiel pour concentrer l'analyse sur les bâtiments commerciaux, industriels et publics.

02

Traitement des valeurs manquantes

Suppression des variables creuses présentant plus de 40% de données manquantes. Imputation du score ENERGY STAR par la médiane du type de propriété correspondant.

03

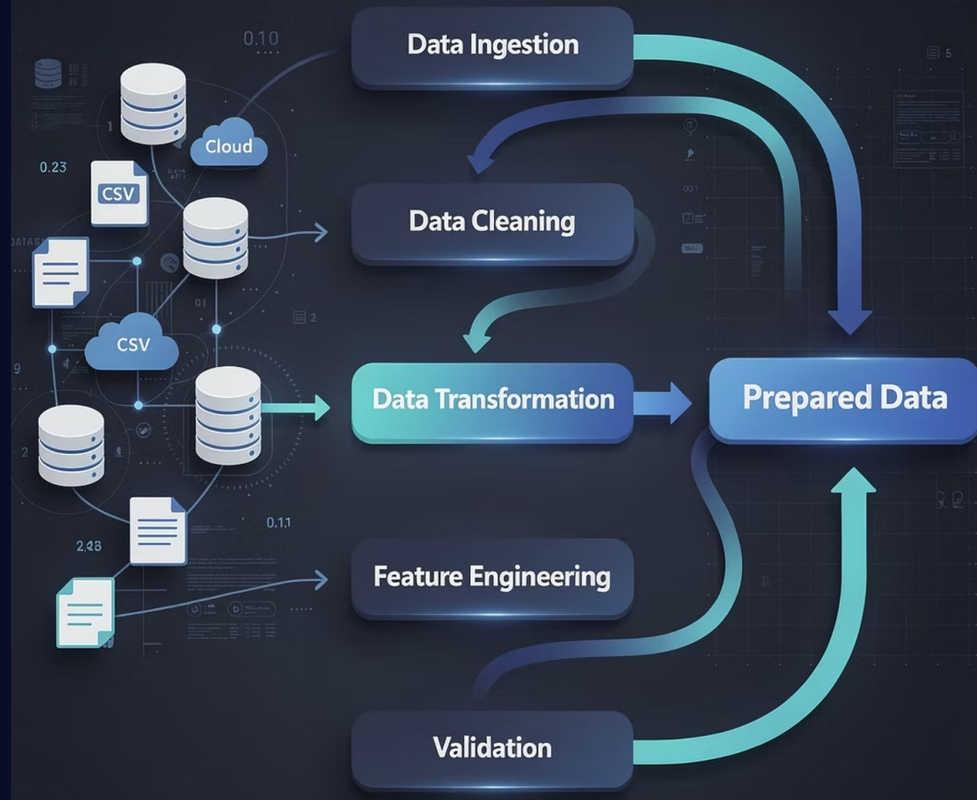
Normalisation de la distribution

Application d'une transformation logarithmique sur la variable cible (SiteEnergyUse) pour stabiliser la variance et obtenir une distribution gaussienne exploitable par les algorithmes.

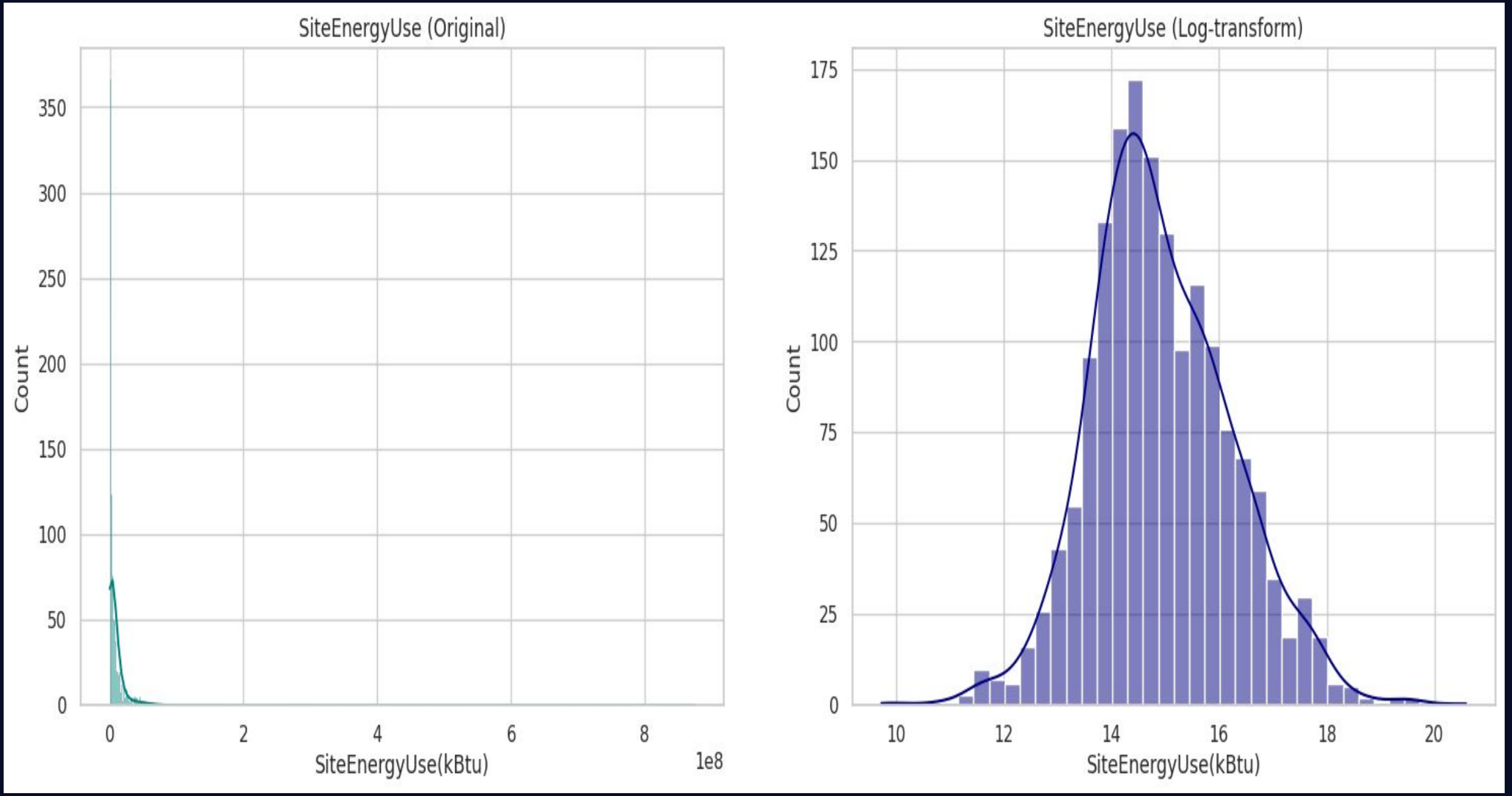
Architecture du dataset final

Variables Structurelles

- PropertyGFATotal : surface totale du bâtiment
- YearBuilt : année de construction
- NumberOfBuildings : nombre de structures
- NumberOfFloors : hauteur en étages
- Latitude & Longitude : coordonnées géographiques...



Distribution de la variable cible avant et après transformation



La consommation brute est très asymétrique. Quelques bâtiments très énergivores étirent la distribution, ce qui pourrait tromper le modèle de régression. La distribution logarithmique se rapproche de la loi normale.

FEATURE ENGINEERING

Enrichissement des variables

Moteurs physiques

Transformation de la variable cible SiteEnergyUse(kBtu) en une forme plus stable pour les algorithmes mathématiques.

Dimension géographique

Création de clusters spatiaux via K-Means pour capturer les effets de densité urbaine. Calcul de la distance au centre-ville (formule de Haversine) pour modéliser l'accessibilité aux réseaux énergétiques.

Dimension temporelle et spatiale

Calcul de l'âge du bâtiment pour intégrer l'évolution des batiments, et création de cluster pour déceler les batiements à grandes surfaces et à surface “normale”

Encodage

Target encoding pour surmonter l’inconvénient du one hot encoding et du binary encoding qui génèrent un nombre important de modalités.

MODÉLISATION

Le Stacking : Architecture à Deux Niveaux

Pour maximiser la précision prédictive (après les modèles baselines), nous avons déployé une architecture de **Stacking**, technique d'ensemble learning combinant plusieurs algorithmes complémentaires.

1

Niveau 1 : Les base learners

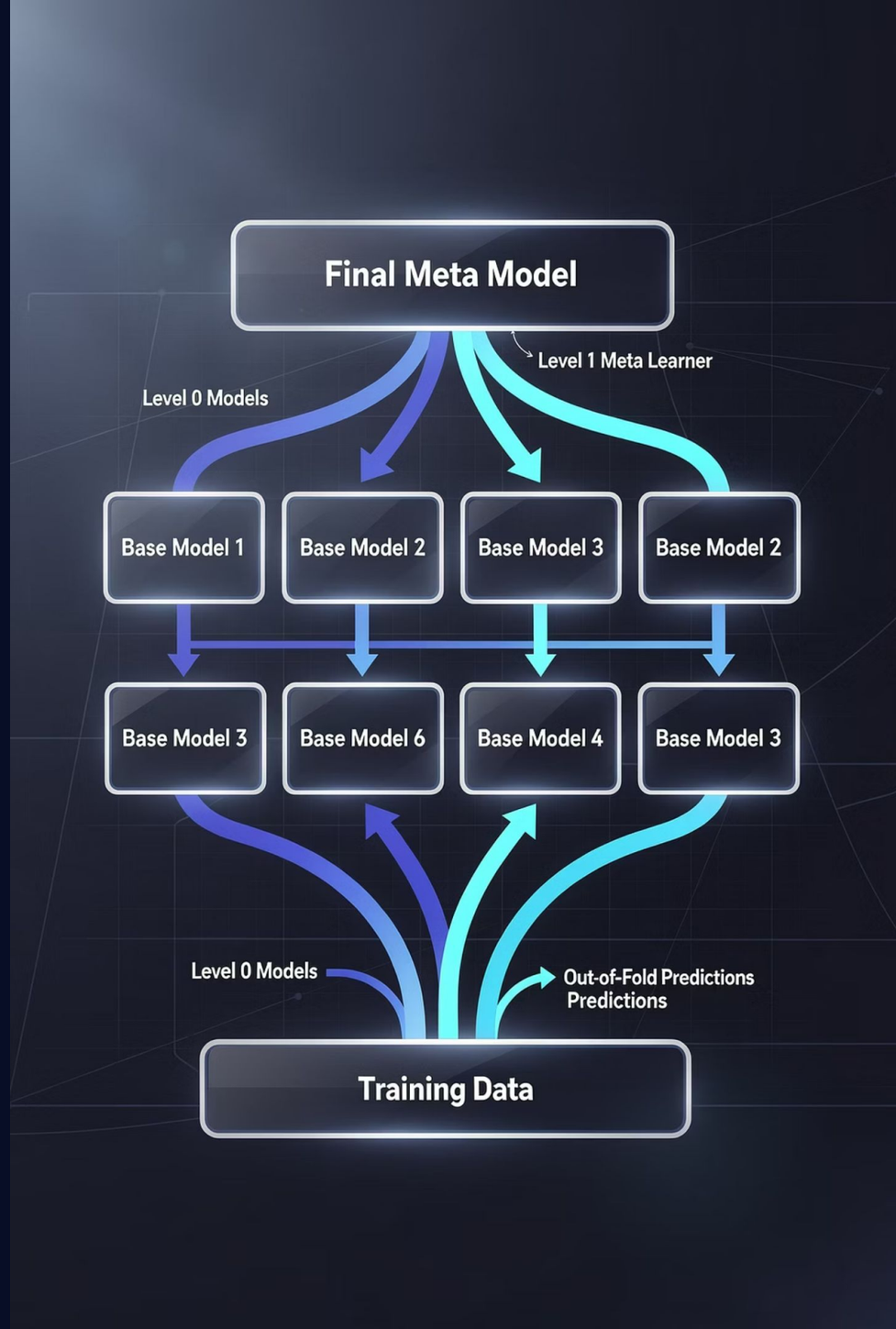
On a utilisé les quatres modèles qui avaient les meilleurs performances lors de l'entraînement :

- ExtraTrees , XGBoost
- RandomForest , SVM

2

Niveau 2 : Le Méta-Apprenant

Plutôt que de faire une simple moyenne, nous utilisons un dernier modèle de Méta-learner qui a aussi donné la meilleur performance parmis plusieurs testés : linear SVR



Les Algorithmes au Cœur du Système

ExtraTrees Regressor

Forêt d'arbres randomisés, pour capturer les interactions non-linéaires entre variables Robuste face au bruit des données.

XGBoost

Algorithme de boosting par gradient, optimisant itérativement les erreurs résiduelles. Particulièrement performant sur les données structurées et tabulaires.

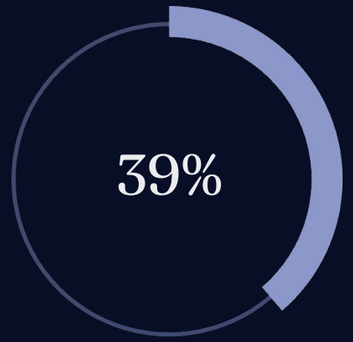
RandomForest

Ensemble d'arbres de décision agrégés, offrant un boncompromis entre précision et interprétabilité(les arbres sont interprétables). Permet de calculer l'importance des features.

Support Vector Machine

Régression par vecteurs de support, traçant des hyperplans optimaux dans l'espace des caractéristiques. Efficace pour modéliser les relations complexes.

Performance du Modèle : Résultats Quantitatifs



MAPE Final

Erreur moyenne absolue en pourcentage

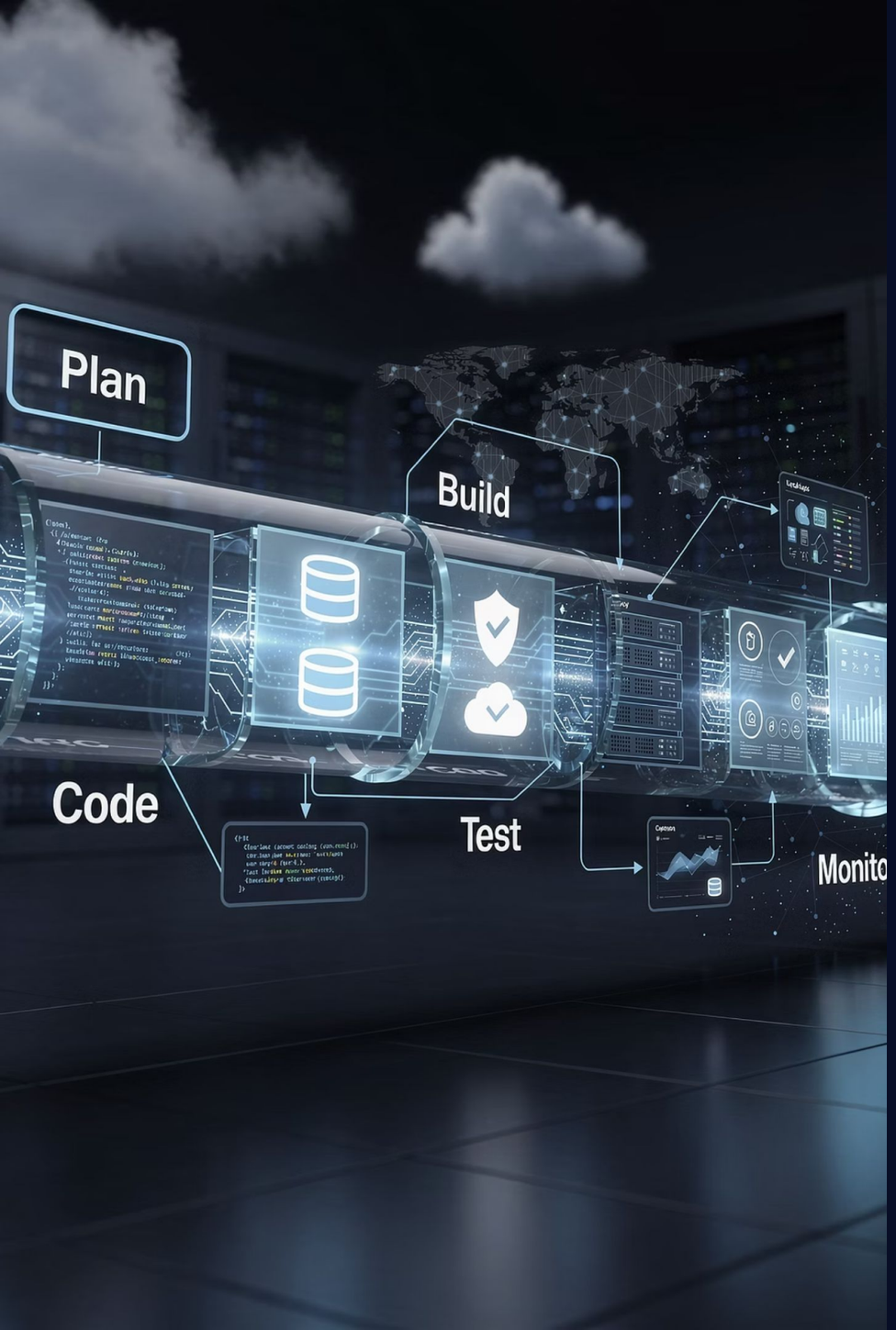
R^2 Score

Coefficient de détermination

Le modèle atteint un **MAPE de 39.%**, ce qui signifie que nos prédictions s'écartent en moyenne de moins de 39% de la consommation réelle. Cette performance est validée par une stratégie de validation croisée K-Fold (k=5), pour mieux garantir la généralisation

Le coefficient R^2 atteint 78 % sur la cible transformée en log, mais s'établit à 52 % une fois les prédictions ramenées aux valeurs réelles. Cet écart s'explique par la sensibilité du R^2 aux échelles de grandeur : en revenant aux données réelles, le poids des très grands bâtiments (outliers) pèse plus lourdement sur la variance, ce qui mécaniquement réduit le score global. Donc le modèle final explique 52% de la variance.

Contexte de Performance : Pour les bâtiments standards (bureaux, commerces), le MAPE descend à 25-30%. Ce sont Les structures complexes type Campus que le modèle ne parvient pas à bien capter en sous estimant leur consommation.



MLOPS

Rigueur industrielle & Cycle de vie

Modularisation

Transition d'un notebook exploratoire vers un code structuré, réutilisable et maintenable avec séparation claire des responsabilités.

Tests Automatisés

Implémentation de tests unitaires et d'intégration pour valider chaque composant du pipeline de données et de modélisation.

Traçabilité MLflow

Suivi exhaustif des expérimentations, hyperparamètres, métriques de performance et artifacts via la plateforme MLflow.

Cette infrastructure MLOps garantit la reproductibilité des résultats, facilite la collaboration entre data scientists et accélère le déploiement en production.

Traçabilité avec MLflow

Tracking des expérimentations

- Versionnage automatique de chaque run d'entraînement
- Comparaison visuelle des hyperparamètres testés
- Historique complet des métriques de performance
- Stockage des artifacts (modèles, graphiques, logs)

Avantages opérationnels

- Reproductibilité totale des résultats scientifiques
- Facilitation des audits et de la gouvernance
- Accélération du débogage et de l'optimisation
- Base de connaissances pour les équipes futures

MLflow devient le journal de bord numérique du projet, documentant chaque décision technique et permettant de revenir à tout moment sur une version antérieure du modèle si nécessaire.

Dashboard

1. OBJECTIF PRINCIPAL

Accompagner la modélisation prédictive de la consommation énergétique.

- Explorer pour mieux modéliser : visualiser et comprendre la structure de la base de données (bâtiments non-résidentiels) avant la conception des algorithmes.
- Aide à la décision : identifier les tendances et la qualité des données pour orienter les choix techniques (nettoyage, transformations).

Lien vers le dashboard : [EnergyML Dashboard | Analyse de Consommation Énergétique]

2. FONCTIONNALITÉS CLÉS

📊 Vue d’ensemble

- Monitoring des indicateurs clés : Nombre de bâtiments, Conso. moyenne, Surface, Score ENERGY STAR.

📈 Analyse univariée

- Compréhension des formes de distribution.
- Détection des outliers et besoin de transformations.

Dashboard

2. FONCTIONNALITÉS CLÉS (SUITE)

📉 Analyse bivariée (Relations)

- Identification des prédicteurs potentiels (Scatter & Box plots).
- Validation des hypothèses de linéarité.

🔗 Matrice de corrélation

- Outil d'aide à la sélection de features.
- Détection de la multicolinéarité.

🚀 Fonctionnalités intelligentes

- Système de multi-filtres et analyses automatiques (recommandations ML).
- Export de rapports d'analyse au format HTML.

La Valeur Stratégique du Score ENERGY STAR

L'analyse confirme que le **score ENERGY STAR** se positionne comme le un prédicteur influent en plus de la surface totale,.

Cette découverte valide l'efficacité des politiques de certification énergétique.

Implications Politiques

- Incitations fiscales renforcées pour les bâtiments certifiés
- Intégration du score dans les critères d'attribution de permis
- Campagnes de sensibilisation ciblées

RECOMMANDATIONS

Plan d'action pour optimiser le système

Priorité 1 : Qualité des données d'entrée

Systématiser la collecte des surfaces exactes (PropertyGFATotal) et des types d'usage détaillés lors du dépôt de permis. Ces deux variables sont les piliers de la précision prédictive.

Priorité 2 : Enrichissement pour campus

Acquérir davantage de données sur les structures complexes multi-usages pour réduire le MAPE à moins de 39% et garantir une couverture uniforme du parc.

Priorité 3 : Promotion de la certification

Multiplier les incitations pour obtenir le score ENERGY STAR dès la conception, créant ainsi un cercle vertueux entre prédiction et performance réelle.

Priorité 4 : Intégration SI existants

Connecter l'API aux systèmes d'information municipaux (GIS, ERP urbanisme) pour automatiser les estimations.

Impact Attendu sur les Politiques Publiques

Pré-diagnostic automatisé

Estimation énergétique lors du dépôt de permis de construire

Évolution réglementaire

Ajustement des normes thermiques basé sur l'analyse des données historiques

Ciblage des aides

Allocation optimale des subventions selon les bâtiments à fort potentiel d'amélioration

Suivi de performance

Comparaison consommation réelle vs. prédite pour identifier les dérives

Conclusion

Performance

- MAPE de 39.72% validé par K-Fold
- Architecture Stacking
- Infrastructure MLOps garantissant la maintenabilité

Valeur opérationnelle

- API
- Dashboard

Impact métier

- Aide à la décision éclairée
- Optimisation de l'allocation des ressources
- Accélération vers l'objectif 2050

Ce projet démontre qu'encore une de plus, les acquis dans le cadre du cours de Machine Learning peuvent constituer un excellent outil d'orientations pour l'atteinte des objectifs de développement durable.