Pour info, sur les slides 9, 16, 22 et 24, il y a une animation qui cache le texte en dessous donc si possible plutôt regarder le ppt en mode diaporama pour voir le texte sur ces slides.

ANTICIPER LES BESOINS EN CONSOMMATION DE BÂTIMENTS

VILLE DE SEATTLE

SOMMAIRE

1. Rappel de la problématique
2. Présentation du jeu de données et manipulations réalisées
3. Approche de modélisation
4. Présentation des résultats
5. Conclusion

1.

Rappel de la problématique

Seattle

Rappel de la problématique

- Objectif : ***ville neutre en émissions de carbone*** en 2050.

- Suivi de la consommation énergétique et des émissions des bâtiments non destinés à l'habitation.

- Relevés sur certains bâtiments en 2015 et 2016.

- ***Relevés coûteux*** → nécessité de ***développer un modèle de prédiction***.

2.

Présentation du jeu de données et manipulation sur les données

Seattle

Présentation du jeu de données

- Relevés effectués par les agents de la ville sur certains bâtiments en 2015 et 2016.

- Données concernant :
  - *Caractéristiques du bâtiment :* type, principale utilisation, année de construction, nombre d'immeubles, nombre d'étages, localisation, surface…
  - *Consommations énergétiques :* consommation du bâtiment (sur l'année et à météo standard), consommation totale incluant les pertes d'acheminement de l'énergie, les émissions de CO_2, la consommation de chaque type d'énergie utilisée…
  - *Autres :* Energy Star Score, commentaires sur les bâtiments…

- 2015 : 3 340 lignes et 47 colonnes
- 2016 : 3 376 lignes et 46 colonnes

Manipulations sur les données

- Regroupement dans un seul jeu de données des relevés de 2015 et 2016 en prenant les relevés de 2016 en cas de doublon.

- Recalcul des données par square foot car pour certaines données :
  - Consommation par square foot ≠ Consommation totale / surface totale

- Filtrage sur les bâtiments non destinés à l'habitation.

- Homogénéisation des catégories pour les variables 'PrimaryPropertyType' et 'Neighborhood'.
  - Exemple : 'Central', 'Ballard' → 'CENTRAL', 'BALLARD'

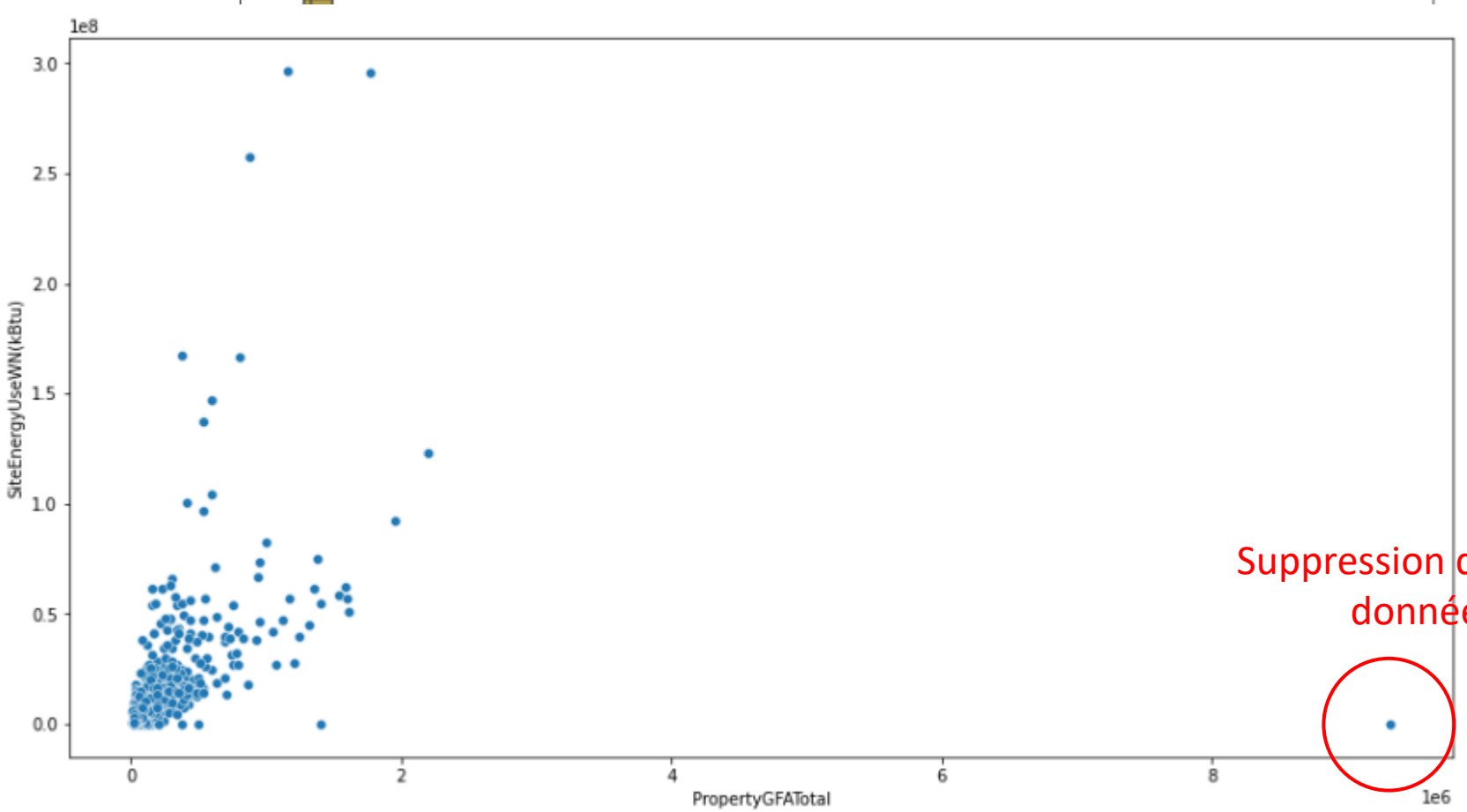
- Pour la variable 'PrimaryPropertyType' :
  - Catégorie 'Office' ajouté à 'Small- and Mid-Sized Office' car que 3 échantillons
  - Catégorie 'Non-Refrigerated Warehouse' ajouté à 'Warehouse' car que 2 échantillons

Manipu

- Analyse de /Type’.
- Analyse de cette ype’.
- Analyse de Total’.
- Suppressio 3tu)’ ou ‘SourceEU 2e)’ et ‘GHGEmiss

Analyse de ‘GHGEmissionsIntensity(kgCO2e/ft2)' en fonction de 'PrimaryPropertyType’ :

Analyse de ‘SiteEnergyUseWN(kBtu)' en fonction de 'PropertyGFATotal’ :



Suppression de cette donnée

Seattle

Manipulations sur les données

- On crée une colonne qui détermine la proportion de surface de parking par rapport à la surface totale.

- On crée des colonnes qui indiquent si le bâtiment utilise tel ou tel type d'énergie (1 si oui et 0 si non).

- Jeu de données final : 1 655 lignes sans valeurs manquantes concernant les caractéristiques, la consommation énergétique ou les émissions de CO2 des bâtiments.

3.
Approche de modélisation

Seattle

Approche de modélisation

- Variables à prédire :
  - SourceEUWN(kBtu) → consommation totale d'énergie
  - GHGEmissions(MetricTonsCO2e) → émissions de CO2

- Test des modèles avec et sans une transformation logarithmique sur chacune des variables à prédire.

- Création de 4 datasets par ordre de complexité pour entrainer les modèles :
  - X1 = ['PrimaryPropertyType', 'NumberofBuildings', 'NumberofFloors', 'PropertyGFATotal']
  - X2 = ['PrimaryPropertyType', 'NumberofBuildings', 'NumberofFloors', 'Neighborhood', 'YearBuilt', 'PropertyGFATotal']
  - X3 = ['PrimaryPropertyType', 'NumberofBuildings', 'NumberofFloors', 'Neighborhood', 'YearBuilt', 'PropertyGFATotal', 'ProportionGFAParking', 'SteamUse', 'NaturalGasUse', 'OtherFuelUse']
  - X4 = ['PrimaryPropertyType', 'NumberofBuildings', 'NumberofFloors', 'Neighborhood', 'YearBuilt', log_PropertyGFATotal', 'ProportionGFAParking', 'SteamUse', 'NaturalGasUse', 'OtherFuelUse']

- Encodage des variables catégorielles avec des OneHotEncoder().
- Test de différentes normalisations des variables numériques :
  - MinMaxScaler()
  - StandardScaler()
  - RobustScaler()

Approche de modélisation

- Métriques utilisées pour évaluer les modèles :
  - Coefficient de détermination (R2)
  - Mean Absolute Error (MAE)
  - Root Mean Squared Error (RMSE)
  - Median Absolute Error (MedAE)

- Création de boucles sur les 4 datasets, sur les différentes normalisations et sur les différents modèles testés.

- Stockage des résultats dans un Dataframe.

Approche de modélisation

- Modèles testés :
  - LinearRegression
  - Lasso
  - Ridge
  - ElasticNet
  - SVR
  - KernelRidge
  - RandomForest
  - XGBoost
  - MLPRegressor

- Dans un 1er temps sans optimisation de paramètres, puis avec une optimisation de paramètres avec une GridSearchCV.

4.
Présentation des résultats

Seattle

Prédiction de la consommation totale d'énergie : résultats de la cross validation

Variable target avec log	Dataset	Modèle	Normalisation	R2	MAE	MedAE	RMSE	Fit time	Score time
No	df1	MLPRegresso							2
No	df1	xgboost							
No	df2	Elastic							
No	df2	KernelRid							
No	df2	Ridge							
No	df2	Lasso							
No	df2	LinearRe							
No	df2	RandomFo							
No	df1	SVR							

Modèle retenu :

SVR avec les paramètres suivants
C=10
epsilon = 0
gamma = 0.0178

Sur le dataset 4 avec une mise à l'échelle des variables numériques en utilisant MinMaxScaler et en faisant un log transform sur la variable target.

Variable target avec log	Dataset	Modèle	Normalisation	R2	MAE	MedAE	RMSE	Fit time	Score time
Yes	df4	SVR							
Yes	df4	xgboos							
Yes	df2	RandomFo							
Yes	df4	Elastic							0,01
Yes	df4	Ridge	RobustScaler	78%	8,0E+06	2,1E+06	2,0E+07	0,02	0,01
Yes	df4	KernelRidge	StandardScaler	78%	8,1E+06	2,1E+06	2,1E+07	0,19	0,03
Yes	df4	Lasso	RobustScaler	77%	8,1E+06	2,1E+06	2,1E+07	0,03	0,01
Yes	df4	LinearReg	MinMaxScaler	77%	8,1E+06	2,2E+06	2,2E+07	0,03	0,01
Yes	df4	MLPRegressor	MinMaxScaler	76%	8,1E+06	2,1E+06	2,2E+07	0,67	0,01

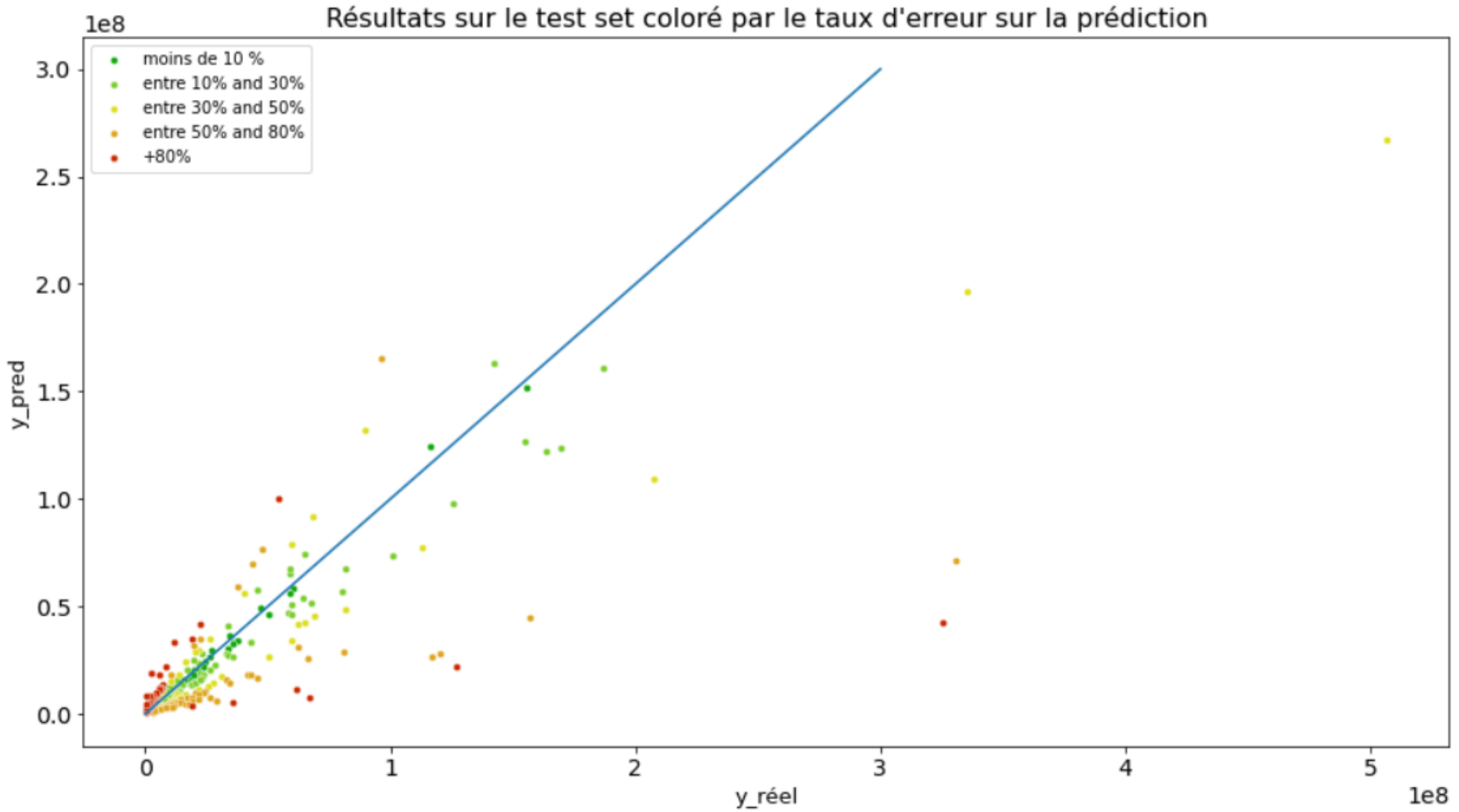
Variable à prédire	SourceEUWN (kBtu)
mean	2,3E+07
std	4,8E+07
min	1,1E+05
25%	3,2E+06
50%	7,0E+06
75%	2,2E+07
max	6,7E+08

Résultat sur le jeu de test

	Variable target avec log	Dataset	Modèle	Normalisation	R2	MAE	MedAE	RMSE
Test set	**Yes**	**df4**	**SVR**	**MinMaxScaler**	**64%**	**9,1E+06**	**2,2E+06**	**2,8E+07**
Cross validation	*Yes*	*df4*	*SVR*	*MinMaxScaler*	*82%*	*7,4E+06*	*2,1E+06*	*1,8E+07*

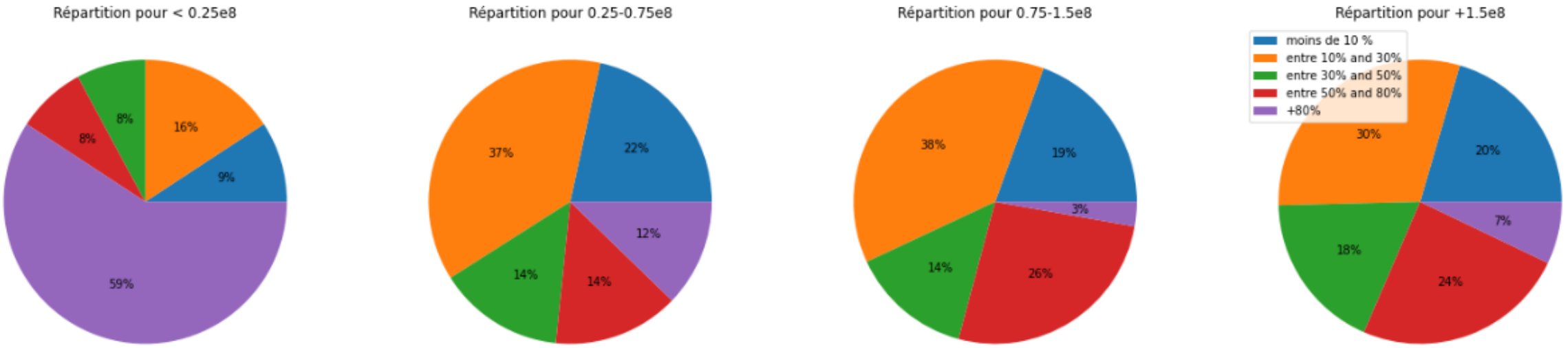
→ Baisse du niveau de performance sur le jeu de test notamment sur la MAE et la RMSE.

Analyse des résultats sur le jeu de test



Analyse des résultats sur le jeu de test

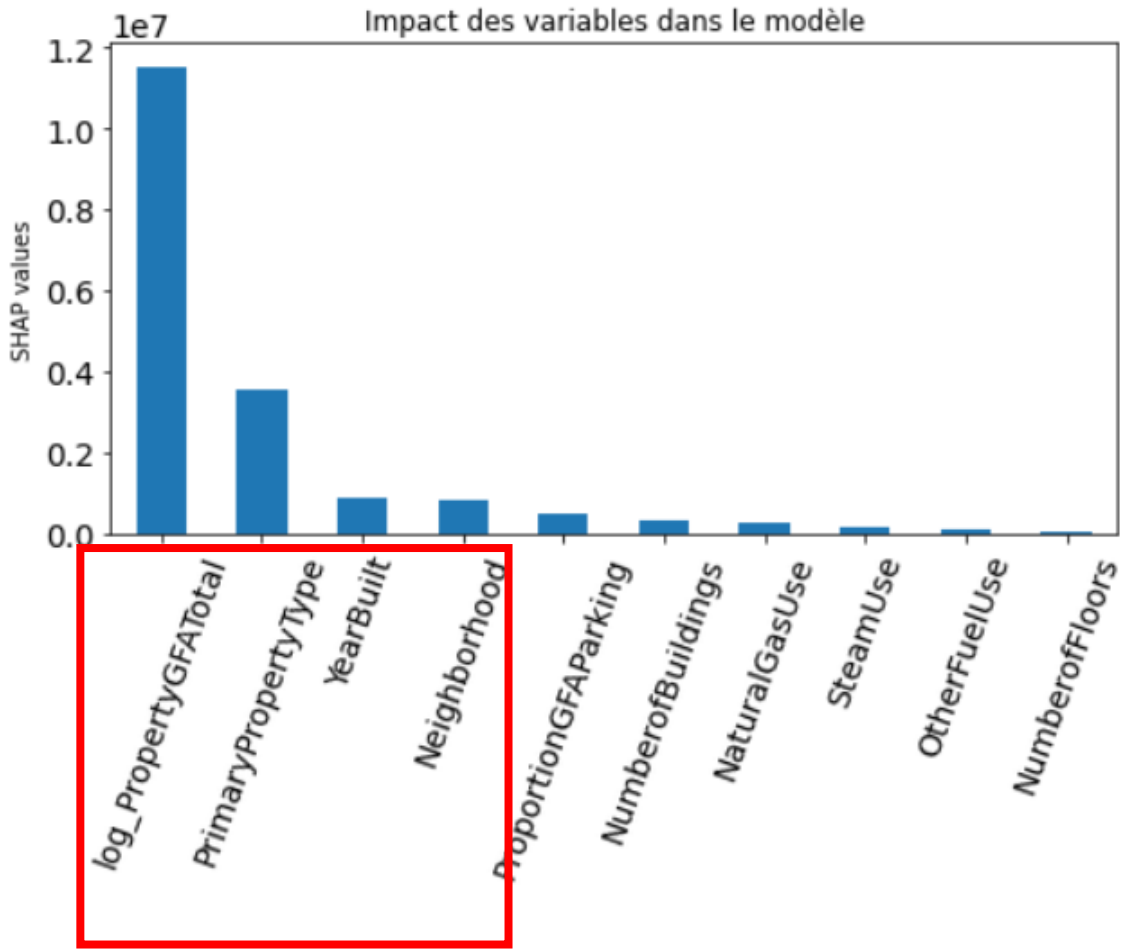
- Répartition du taux d'erreur sur le jeu de test en fonction de la valeur de la variable target :



→ On retrouve bien un plus fort taux d'erreur pour les petites valeurs de jeu de test

Impact des variables sur le résultat de la prédiction

- Utilisation de SHAP en calculant les SHAP values



Seattle

Impact de l'ENERGY STAR Score sur la prédiction

- Données manquantes → réduction du jeu de données à 1 096 échantillons (vs 1 655 = -34%)

- Entraînement du modèle retenu sur 2 datasets :
  - Dataset retenu lors de la sélection du modèle
  - Dataset retenu lors de la sélection du modèle en ajoutant l'ENERGY STAR Score

- Comparaison des résultats obtenus après cross validation :

ENERGY STAR Score	Dataset	Modèle	Normalisation	R2	MAE	MedAE	RMSE	Fit time	Score time
SANS	df4	SVR	MinMaxScaler	79%	7,41E+06	2,23E+06	2,14E+07	0,12	0,01
AVEC	df5	SVR	MinMaxScaler	72%	5,34E+06	1,35E+06	1,89E+07	0,08	0,01

→ Baisse du r2 mais amélioration significative des autres métriques.

Prédiction des émissions de CO2 : résultats de la cross validation

Variable target avec log	Dataset	Modèle	Normalisation	R2	MAE	MedAE	RMSE	Fit time	Score time
No	df4	MLPRegress							02
No	df4	RandomF							
No	df3	Elasti							
No	df3	Ridge							
No	df3	KernelRi							
No	df3	Lasso							
No	df3	LinearR							
No	df3	xgboos							
No	df3	SVR							

Variable target avec log	Dataset	Modèle	Normalisation	R2	MAE	MedAE	RMSE	Fit time	Score time
Yes	df3	xgboos							
Yes	df4	SVR							
Yes	df4	RandomFo							
Yes	df4	KernelRidge							0,02
Yes	df4	Ridge	RobustScaler	54%	101	23	345	0,03	0,01
Yes	df4	Elastic	StandardScaler	52%	102	23	358	0,35	0,01
Yes	df4	MLPRegressor	StandardScaler	51%	95	21	397	1,71	0,01
Yes	df4	Lasso	MinMaxScaler	35%	109	23	495	0,02	0,01
Yes	df4	LinearReg	MinMaxScaler	-121%	134	23	1123	0,03	0,01

Modèle retenu :

XGBoost avec les paramètres suivants
n_estimators=15
min_child_weight = 1
max_depth = 6
learning_rate = 0.5
gamma = 0.25

Sur le dataset 3 avec une mise à l'échelle des variables numériques en utilisant RobustScaler et en faisant un log transform sur la variable target.

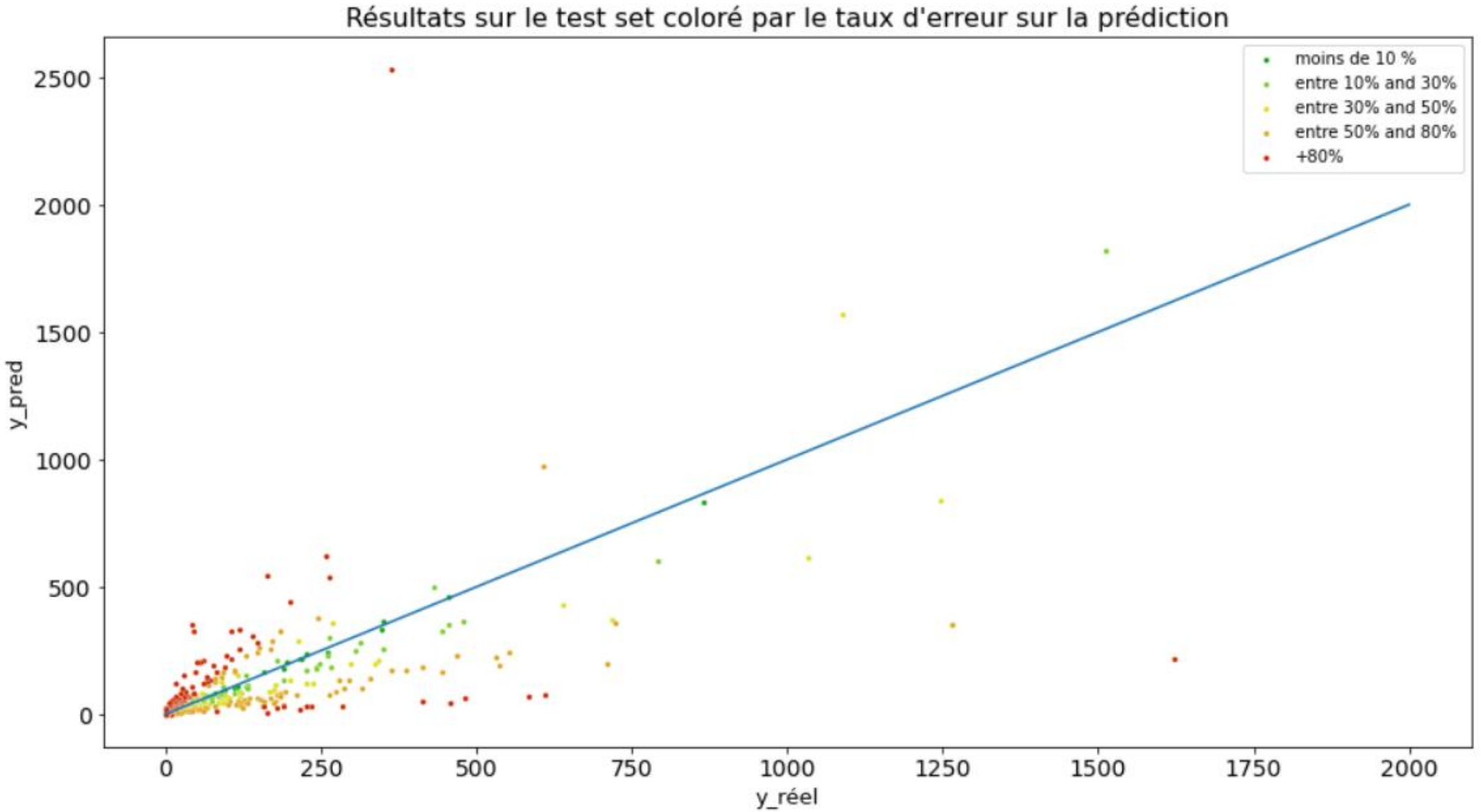
Variable à prédire	GHGEmissions (MetricTonsCO2e)
mean	161,3
std	594,6
min	0,6
25%	20,9
50%	49,6
75%	139,4
max	12307,2

Résultat sur le jeu de test

	Variable target avec log	Dataset	Modèle	Normalisation	R2	MAE	MedAE	RMSE
Test set	**Yes**	**df3**	**xgboost**	**RobustScaler**	**68%**	**91**	**24**	**296**
Cross validation	*Yes*	*df3*	*xgboost*	*RobustScaler*	*65%*	*88*	*22*	*287*

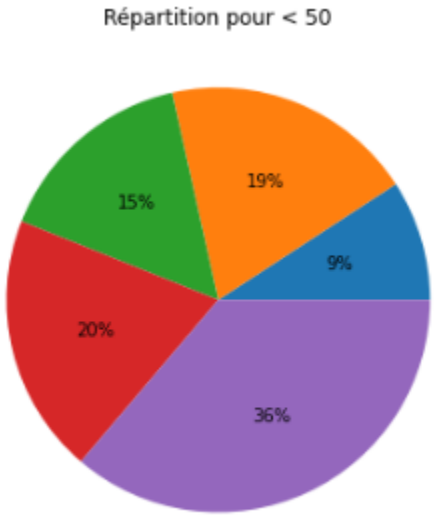
→ Niveau de performance équivalent avec une légère hausse du r2 mais des autres métriques légèrement dégradées.

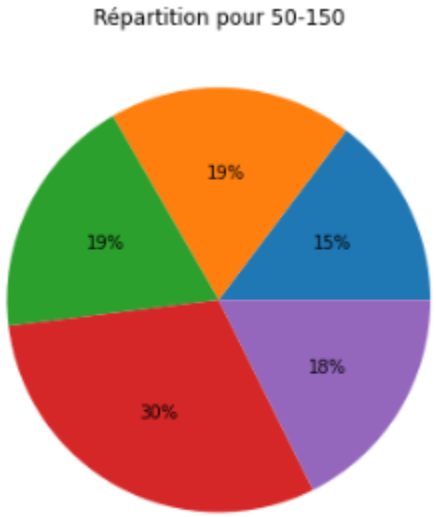
Analyse des résultats sur le jeu de test

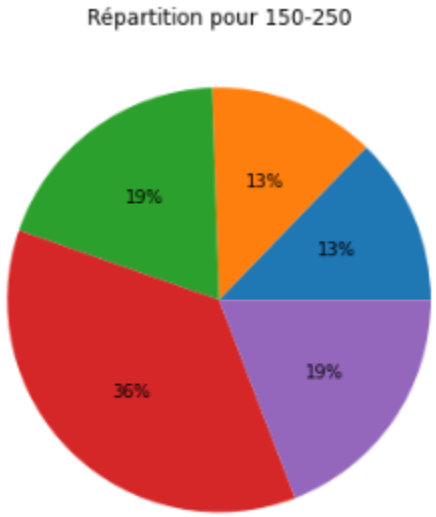


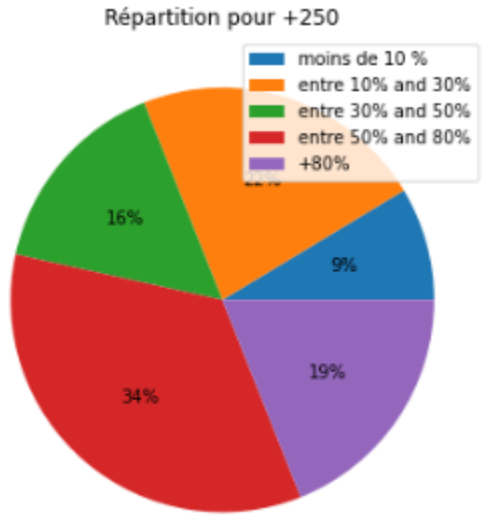
Analyse des résultats sur le jeu de test

- Répartition du taux d'erreur sur le jeu de test en fonction de la valeur de la variable target :





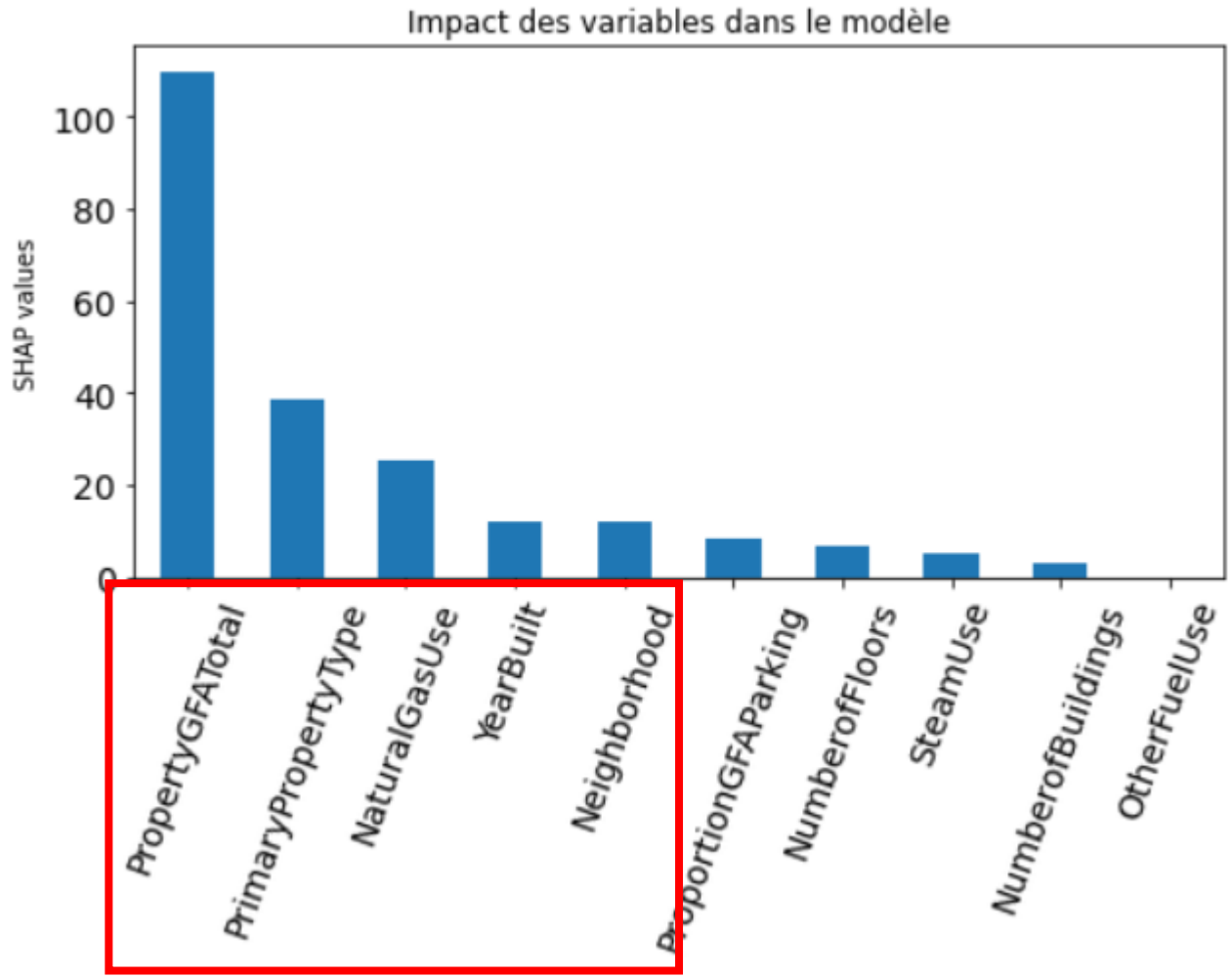




→ On retrouve bien un plus fort taux d'erreur pour les petites valeurs de jeu de test

Impact des variables sur le résultat de la prédiction

- Utilisation de SHAP en calculant les SHAP values



Seattle

Impact de l'ENERGY STAR Score sur la prédiction

- Données manquantes → réduction du jeu de données à 1 096 échantillons (vs 1 655 = -34%)

- Entraînement du modèle retenu sur 2 datasets :
  - Dataset retenu lors de la sélection du modèle
  - Dataset retenu lors de la sélection du modèle en ajoutant l'ENERGY STAR Score

- Comparaison des résultats obtenus après cross validation :

ENERGY STAR Score	Dataset	Modèle	Normalisation	R2	MAE	MedAE	RMSE	Fit time	Score time
SANS	df3	XGBoost	RobustScaler	56%	88	24	385	0,07	0,02
AVEC	df5	XGBoost	RobustScaler	61%	81	20	297	0,10	0,01

→ Amélioration significative des toutes les métriques.

5.
Conclusion

Seattle

Conclusion

- Deux modèles différents retenus pour la prédiction de chacune des variables
- Application du logarithme sur la variable target pour les deux modèles
- Plus de difficultés à prédire les petites valeurs pour les deux modèles
- Impact important des variables suivantes pour les deux modèles (par ordre d'importance) :
  - PropertyGFATotal (= Surperficie du bâtiment)
  - PrimaryPropertyType (= Utilisation principale du bâtiment)
  - YearBuilt (= Année de construction)
  - Neighborhood (= Quartier)
- Amélioration des métriques lors de l'ajout de l'ERNERGY STAR Score pour les deux modèles mais entraînement du modèle sur 34% de données en moins.