# COMMENT PRÉDIRE LES PRIX DE BIENS IMMOBILIERS ?

CHALLENGEDATA

# PRÉSENTATION DU JEU DE DONNÉES

| | id_annonce | price |
|---|---|---|
| 0 | 35996577 | 355000.0 |
| 1 | 35811033 | 190000.0 |
| 2 | 35731841 | 39000.0 |
| ... | ... | ... |
| 37365 | 36010245 | 328000.0 |
| 37366 | 35864579 | 463000.0 |
| 37367 | 35748883 | 69930.0 |

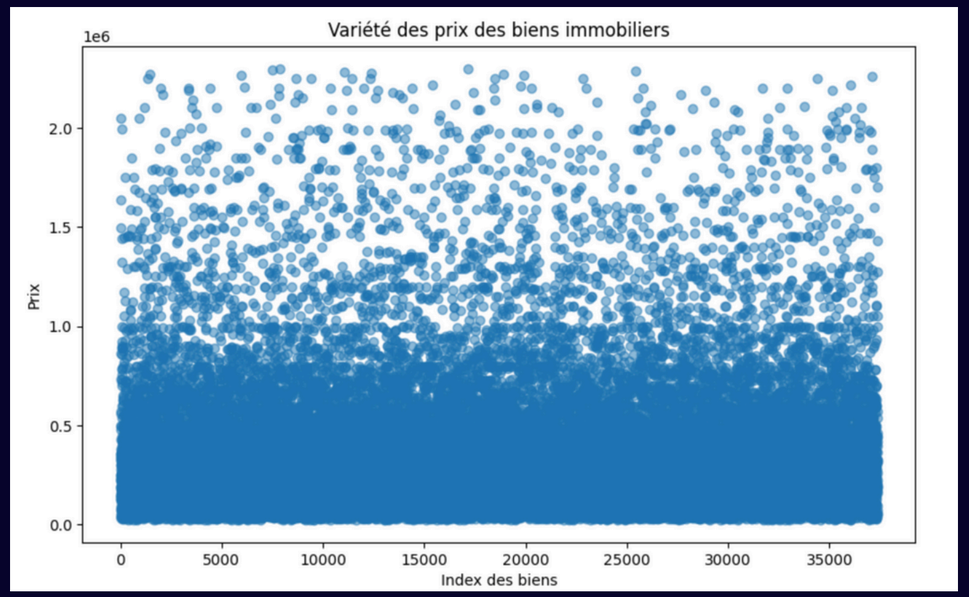
| | id_annonce | property_type | approximate_latitude | approximate_longitude | ... | has_a_garage | has_air_conditioning | last_floor | upper_floors |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 35996577 | appartement | 43.643880 | 7.117183 | ... | 0.0 | 0.0 | 0.0 | 0.0 |
| 1 | 35811033 | appartement | 45.695757 | 4.895610 | ... | 0.0 | 0.0 | 0.0 | 0.0 |
| 2 | 35731841 | maison | 47.966791 | -1.220451 | ... | 0.0 | 0.0 | 0.0 | 0.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 37365 | 36010245 | appartement | 44.397837 | -1.164312 | ... | 0.0 | 0.0 | 0.0 | 0.0 |
| 37366 | 35864579 | duplex | 48.864204 | 2.808693 | ... | 0.0 | 0.0 | 0.0 | 0.0 |
| 37367 | 35748883 | maison | 46.032306 | 1.966711 | ... | 0.0 | 0.0 | 0.0 | 0.0 |

- **TARGET VARIABLE:** *PRICE*
→ DISTRIBUTION DISPERSÉE

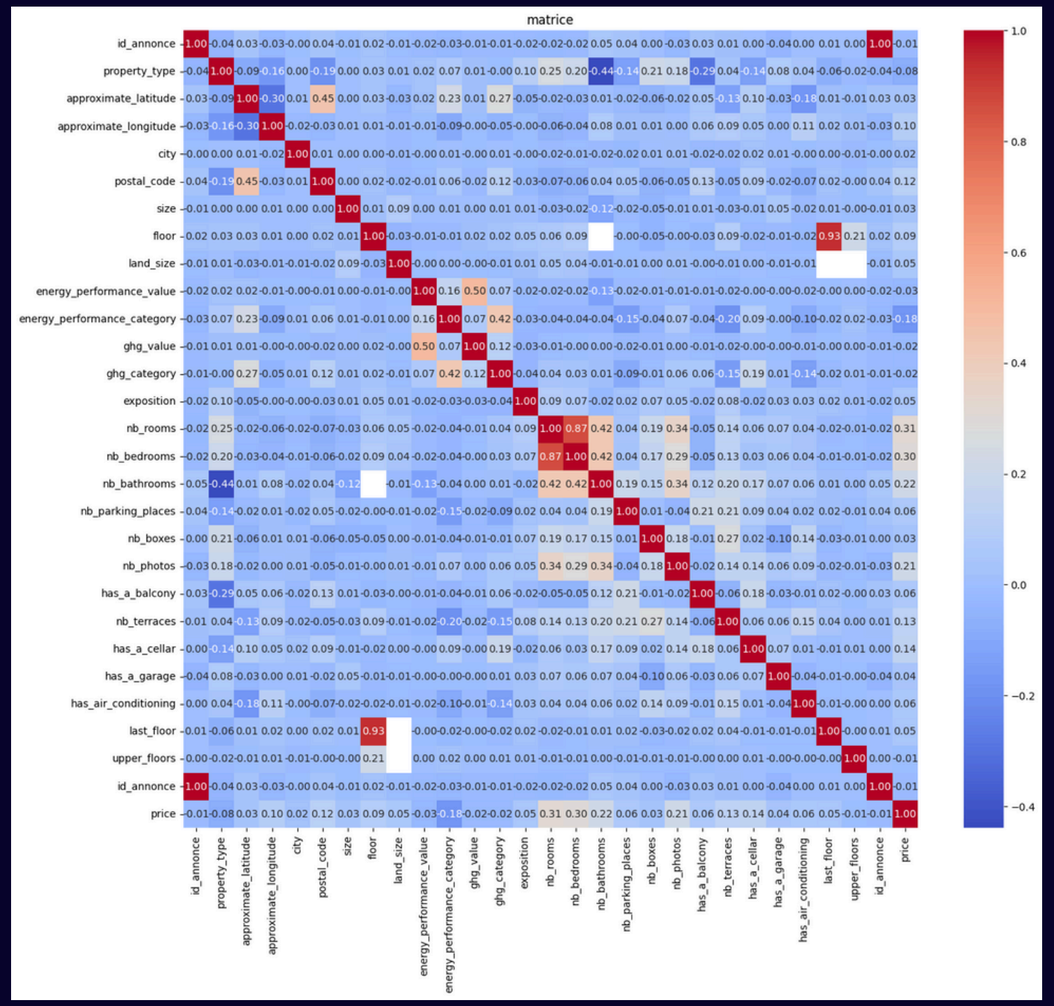
- **VARIABLE D'ENTRÉE:** *ID_ANNONCE*, *CITY*...
→ FEATURES CARACTÉRISANT LES BIENS IMMOBILIERS

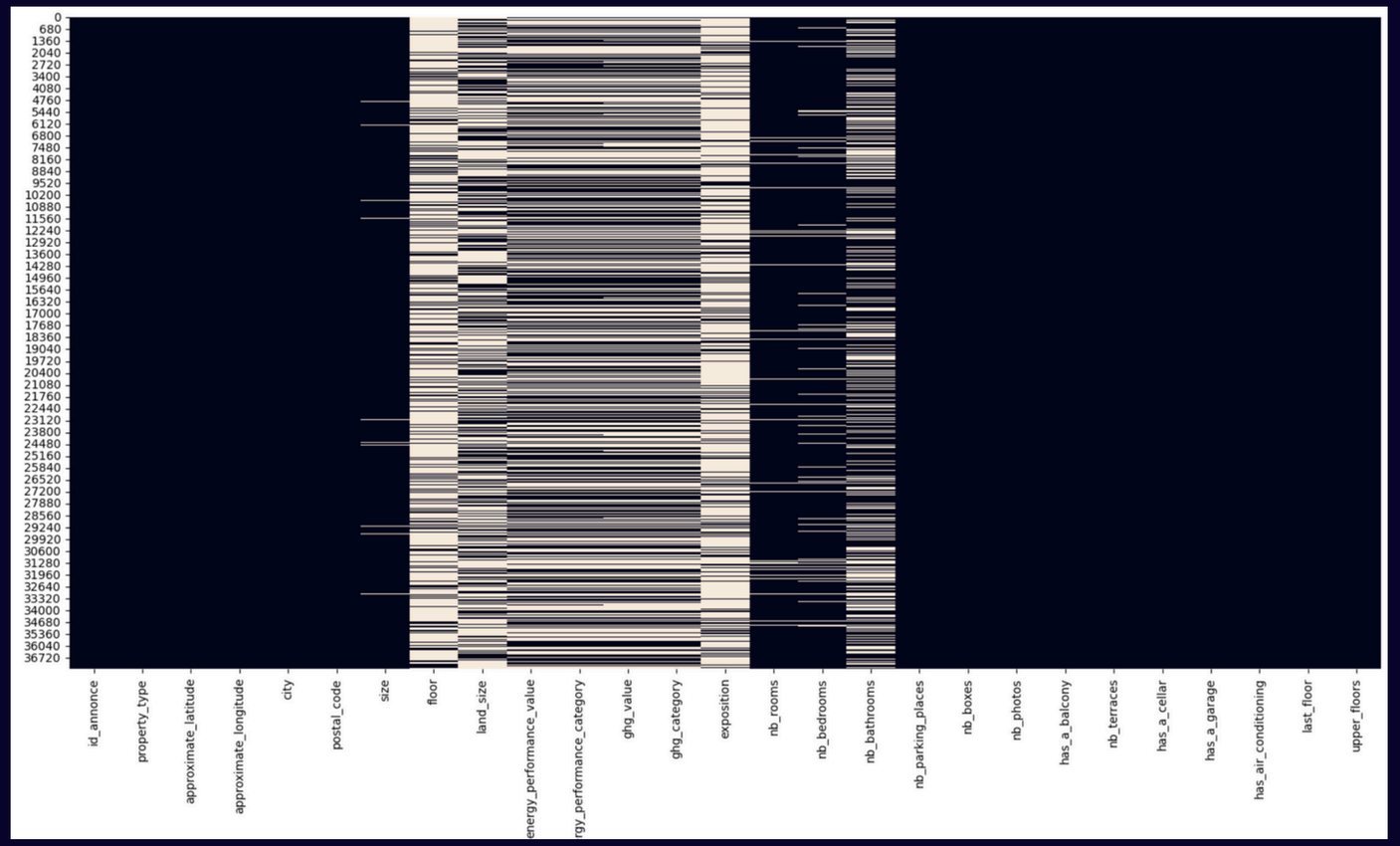
- **DATASET VOLUMINEUX**: 40 000 LIGNES ET 27 COLONNES.

- **VARIABLES NUMÉRIQUES VS CATÉGORIELLES**



# PRÉSENTATION DU JEU DE DONNÉES



# PRÉTRAITEMENT DES DONNÉES

| VARIABLES CATÉGORIELLES | DÉFAUTS | |
|---|---|---|
| ~~CITY~~ | 8000 VALEURS UNIQUES + REDONDANCE | |
| ~~GHG_CATEGORY~~ | 18300 VALEURS MANQUANTES + REDONDANCE | |
| ~~ENERGY_PERFORMANCE_CATEGORY~~ | 18838 VALEURS MANQUANTES + REDONDANCE | |
| ~~EXPOSITION~~ | 75% VALEURS MANQUANTES | |
| PROPERTY_TYPE | LABEL ENCODING | ONE HOT ENCODING |

# PRÉTRAITEMENT DES DONNÉES

| | |
|---|---|
| SIZE | 512 |
| ~~FLOOR~~ | ~~27625~~ |
| LAND_SIZE | 21787 |
| ENERGY_PERFORMANCE_VALUE | 18300 |
| GHG_VALUE | 18838 |
| NB_ROOMS | 1566 |
| NB_BEDROOMS | 2733 |
| NB_BATHROOMS | 13273 |

**IMPUTATION DES NAN**

**K-NEAREST-NEIGHBORS IMPUTER**

- IDENTIFICATION DES K OBSERVATIONS LES PLUS PROCHES
- ESTIMATION PAR LA MOYENNE DES VALEURS

# III) CHOIX DU MODÈLE

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|A_i - F_i|}{A_i}$$

$A_i$ is the actual value

$F_i$ is the forecast value

n is total number of observations

**<-- Pour optimiser :**

-Logistic Regression ?

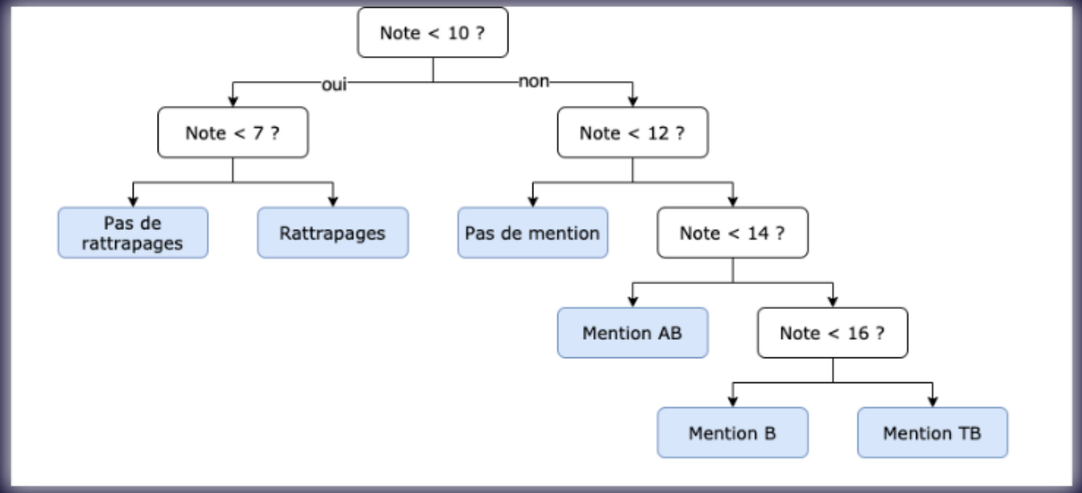
-KNN ?

-Random Forest ?

# IV) OPTIMISATION ET RÉSULTATS

## A : XGBOOST

### Arbres de décisions




### Fonction de perte

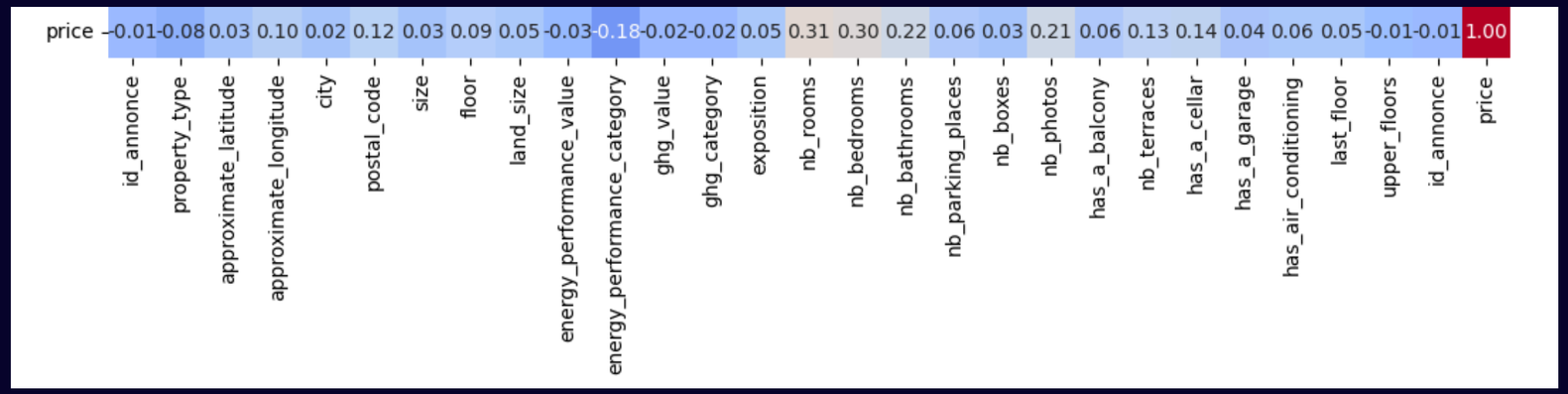
$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

### Différences avec RF :

- BOOSTING / BAGGING
- SÉQUENTIELLEMENT / INDÉPENDEMMENT
- PLUS PERFORMANT SUR DE GRANDS JEUX DE DONNÉES

# IV) OPTIMISATION ET RÉSULTATS

## B : RÉSULTATS




## NOS RÉSULTATS

- 34% - 37%
- 29%


ITERATIVE IMPUTER ?
SIMPLE IMPUTER ?

## MEILLEURS HYPERPARAMÈTRES

{'SUBSAMPLE': 0.91, 'SCALE_POS_WEIGHT': 6, 'REG_LAMBDA': 0.00015, 'REG_ALPHA': 0.005, 'N_ESTIMATORS': 1600, 'MIN_CHILD_WEIGHT': 4, 'MAX_DEPTH': 10, 'LEARNING_RATE': 0.04, 'GAMMA': 0.11, 'COLSAMPLE_BYTREE': 0.85}

# V) IMAGES

| mean_pixel_value | std_pixel_value | mean_width | mean_height |
|---|---|---|---|
| 146.552336 | 49.782720 | 150.75 | 102.000000 |
| 106.383582 | 54.584047 | 228.00 | 171.000000 |
| 122.392555 | 46.708905 | 114.00 | 64.000000 |
| 127.150122 | 58.966517 | 107.00 | 90.333333 |
| 178.711864 | 48.242183 | 257.00 | 171.000000 |
| ... | ... | ... | ... |
| 196.897449 | 45.979444 | 146.00 | 97.000000 |
| 173.689712 | 48.802681 | 257.00 | 171.000000 |
| 167.519529 | 40.034761 | 97.20 | 70.200000 |
| 155.557189 | 49.343631 | 87.00 | 54.200000 |

Le plus corrélée

## Performance améliorée ?

NON, score de 31.4 aprés optimisation

## Pour aller plus loin

-Utiliser des modéles pré-entrainés pour extraire des features plus complexes
-Utiliser des techniques de deep learning