Estimateur du prix des biens immobiliers : combien vaut votre bien ?

TRAN Marilyn
VIGNESWARAN Tharsiya

Dataset

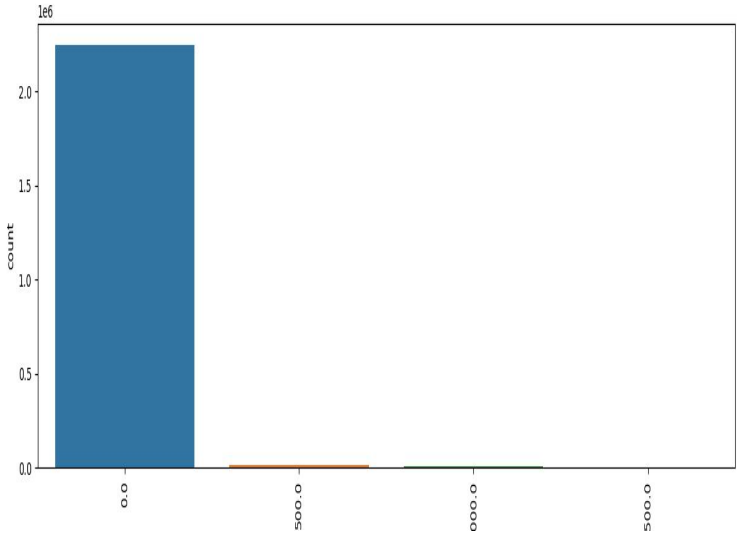
- Source : Valeurs foncieres 2020-2023
  www.data.gouv.fr/fr/datasets/demandes-de-valeurs-foncieres-geolocalisees/

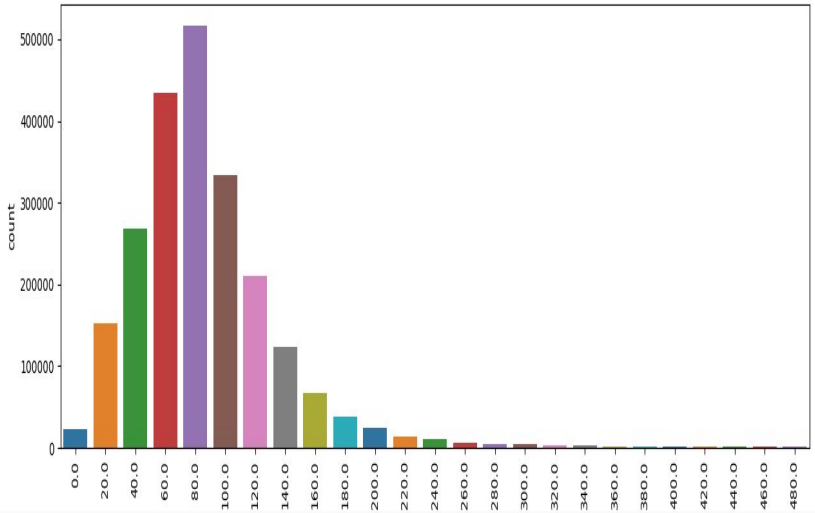
- Colonnes importantes : nature_mutation (Vente/Vente futur), valeur_fonciere, adresse complète avec code postal, type_local (Maison Appartement), surface_reelle_bati, longitude/latitude

=> totale de donnée : +5M

Analyse du Dataset

- Colonnes manquantes : prix du mètre carré ? region ?
- Peu de bien au dessus de 500m^2 et disparité





Nettoyage du Dataset : Les valeurs Null

- Types de mutation “Vente” et “Vente en l’état futur d’achèvement” pour éviter n’inclure “ventes de terrains / ventes issus d’adjudication”

- Nous ne prenons que les ventes dont le nombre de m² bâti est renseigné et différent de zéro

- Nous ne prenons que les ventes dont le prix de ventes est renseigné et différent de zéro

- Nous enlèverons les biens dont l’adresse n’est pas indiquée

Encoding : One Hot Encoding

Region : “Auvergne-Rhone-Alpes” “IDF” etc …

=> region_Auvergne-Rhone-Alpes, region_IDF etc …

-> Ce qui nous donne 13 nouvelles colonnes

Même principe pour :

- nature_mutation (Vente/Vente futur)
- type_local (Maison/Appartement)

=> nombre total de données après nettoyage : ~2M

Preprocessing : Prix du mètre carré en fonction du voisinage

On met l'hypothèse que le prix d'un bien est fortement lié au prix de son voisinage :

- BallTree

  Au lieu de créer un modèle pour les 5 millions de lignes, un modèle BallTree sera créé par région, puis pour chaque ligne, on lui appliquera le modèle de sa région.

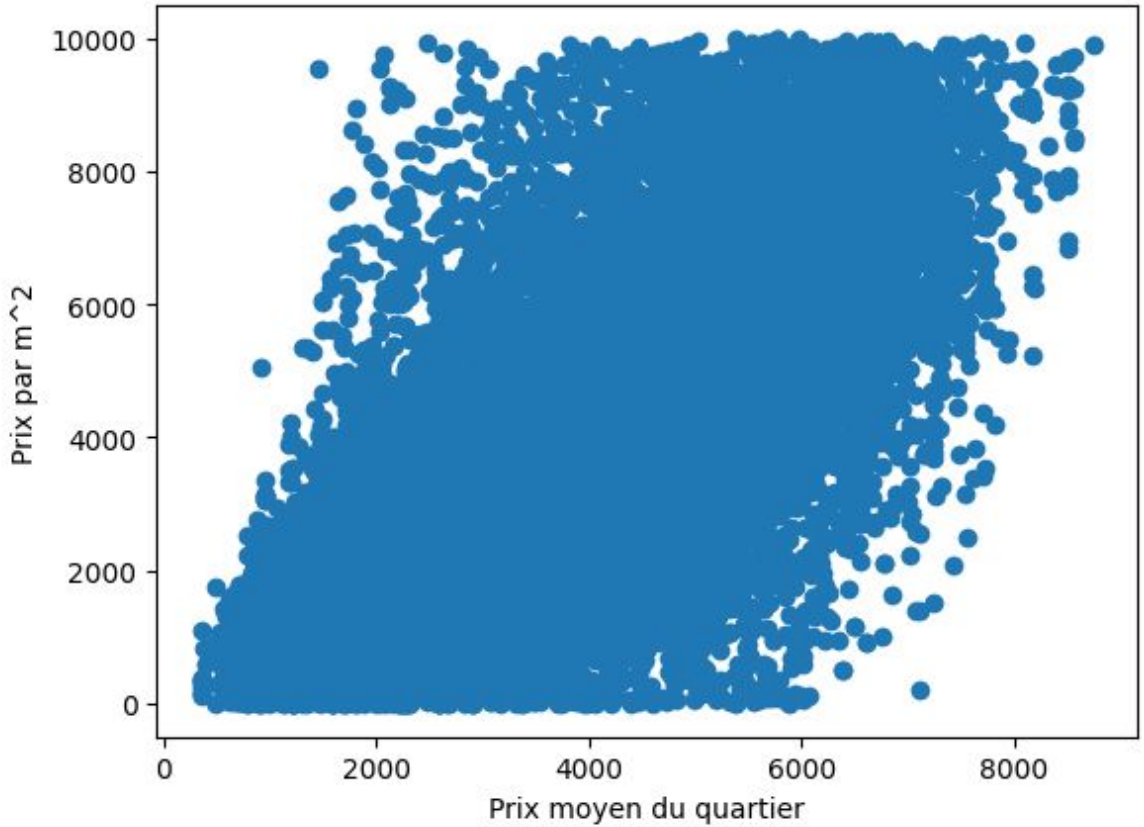
Mais cela est-il suffisant …

➔ prix du bien selon les biens voisins de même type
➔ prix du bien selon les biens voisins similaire en surface
➔ combinaison des deux idées

nature_mutation	Vente
type_local	
Appartement	378202
Local industriel. commercial ou assimilé	101199
Maison	1770303

Résultats

```
[8264 rows x 17 columns]
```



Tuning : remplissage des valeurs Null

Utilisation de la médiane pour remplir les colonnes :

- surface_reelle_batie : médiane avec les K voisins du bien
- valeur_fonciere : médiane avec les K voisins du bien

Train / Test

- Split_train_test :

Features: [type local,surface réelle bati, latitude, longitude,région,prix moyen du quartier]

-> Output : prix du bien par m2

Algo / Modele

- Random Forest Regressor

Evaluation du modèle

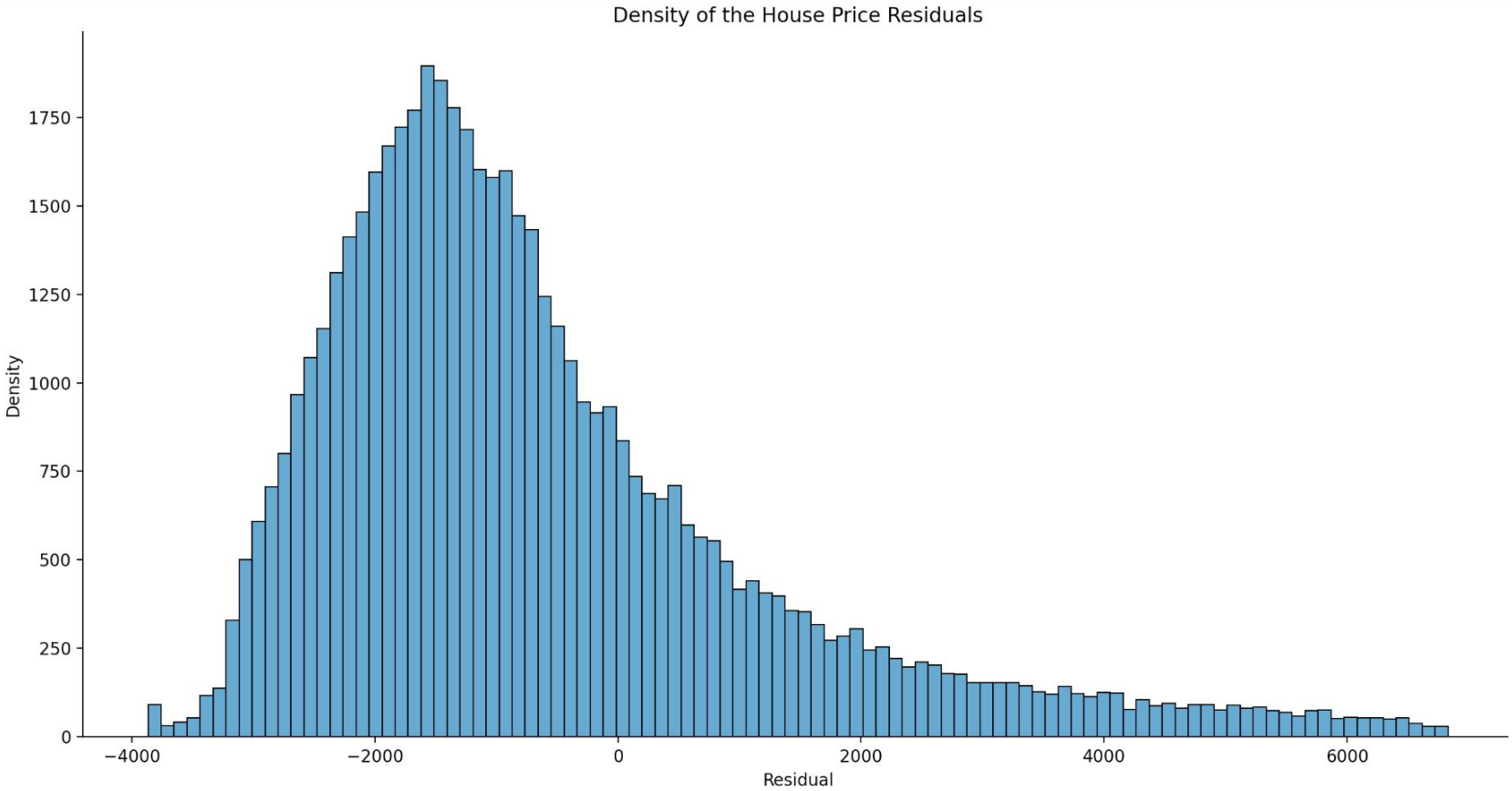
- Le prix par m2 en fonction du prix du voisinage n’est pas totalement linéaire

  => Il n’y a pas de corrélation entre les deux ?

Alors comment estimer son bien ?

Random Forest Regression

Random Forest Regression

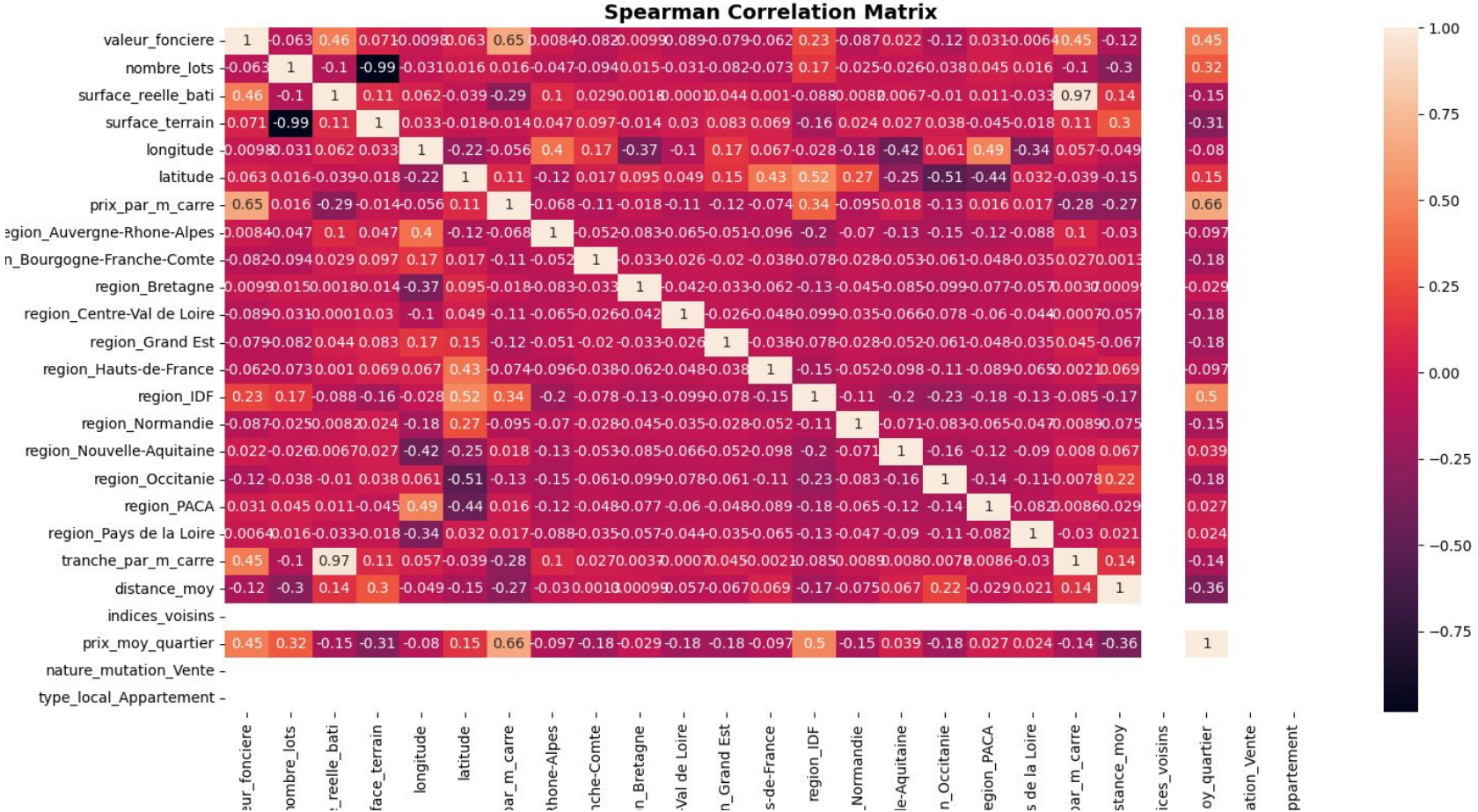


Evaluation d’un autre modèle

Après ces résultats, nous nous sommes penchées sur un autre modèle :

Polynomial Regression

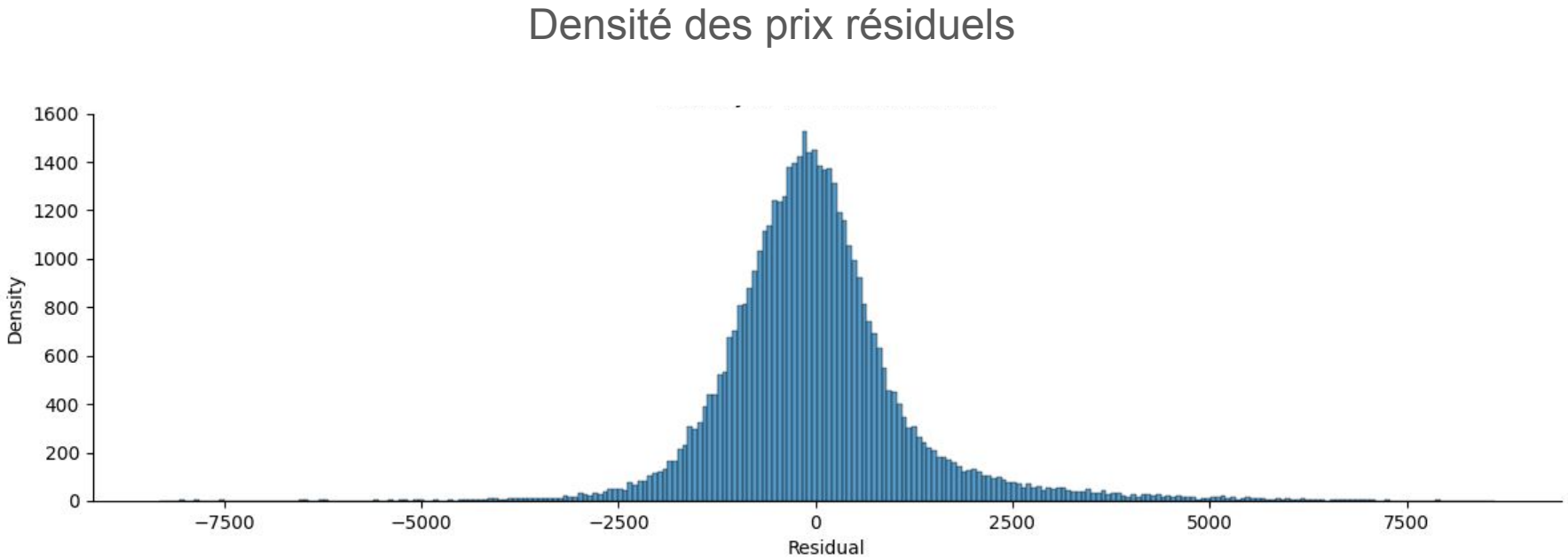
Pour prendre en compte la non linéarité des prix des biens immobiliers


Spearman Correlation Matrix
valeur_fonciere
nombre_lots
surface_reelle_bati
surface_terrain
longitude
latitude
prix_par_m_carre
region_Auvergne-Rhone-Alpes
region_Bourgogne-Franche-Comte
region_Bretagne
region_Centre-Val de Loire
region_Grand Est
region_Hauts-de-France
region_IDF
region_Normandie
region_Nouvelle-Aquitaine
region_Occitanie
region_PACA
region_Pays de la Loire
tranche_par_m_carre
distance_moy
indices_voisins
prix_moy_quartier
nature_mutation_Vente
type_local_Appartement
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Polynomial Regression (séparation par type de biens)


```
Mean Absolute Error (MAE): 828.18
R-squared (R2): 0.55
```

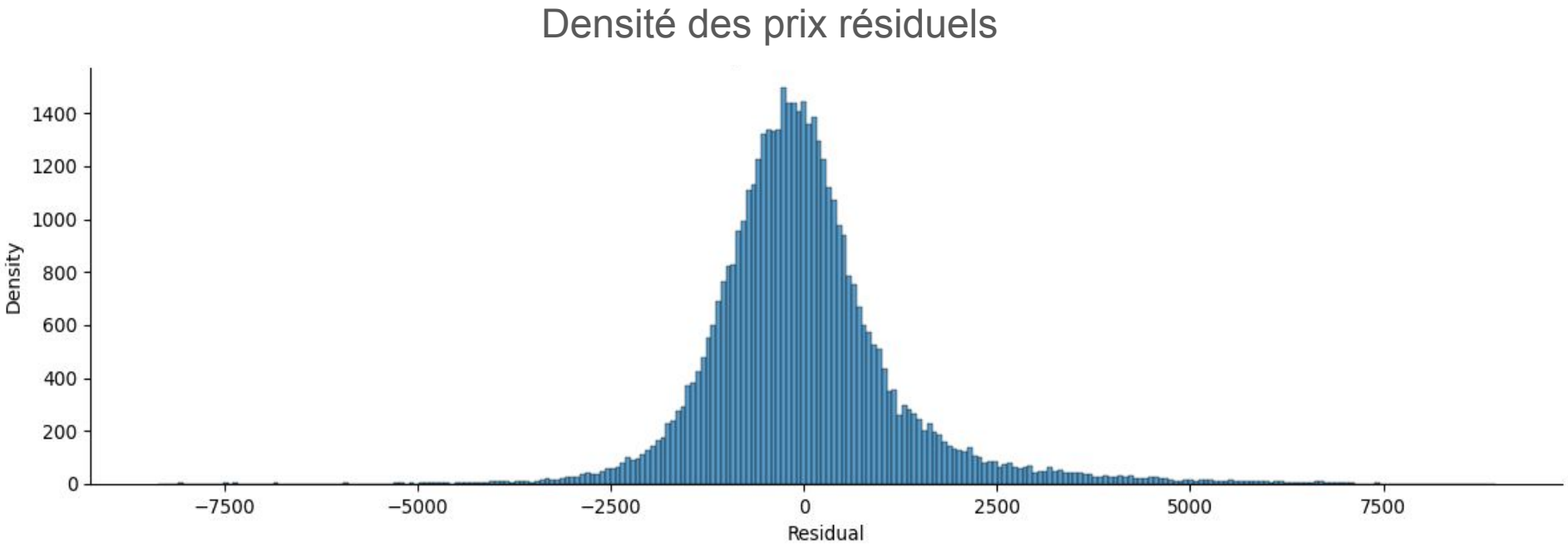
Polynomial Regression (séparation par type de biens)



Polynomial Regression (séparation type et tranche de m carré)


```
Mean Absolute Error (MAE): 860.61
R-squared (R2): 0.52
```

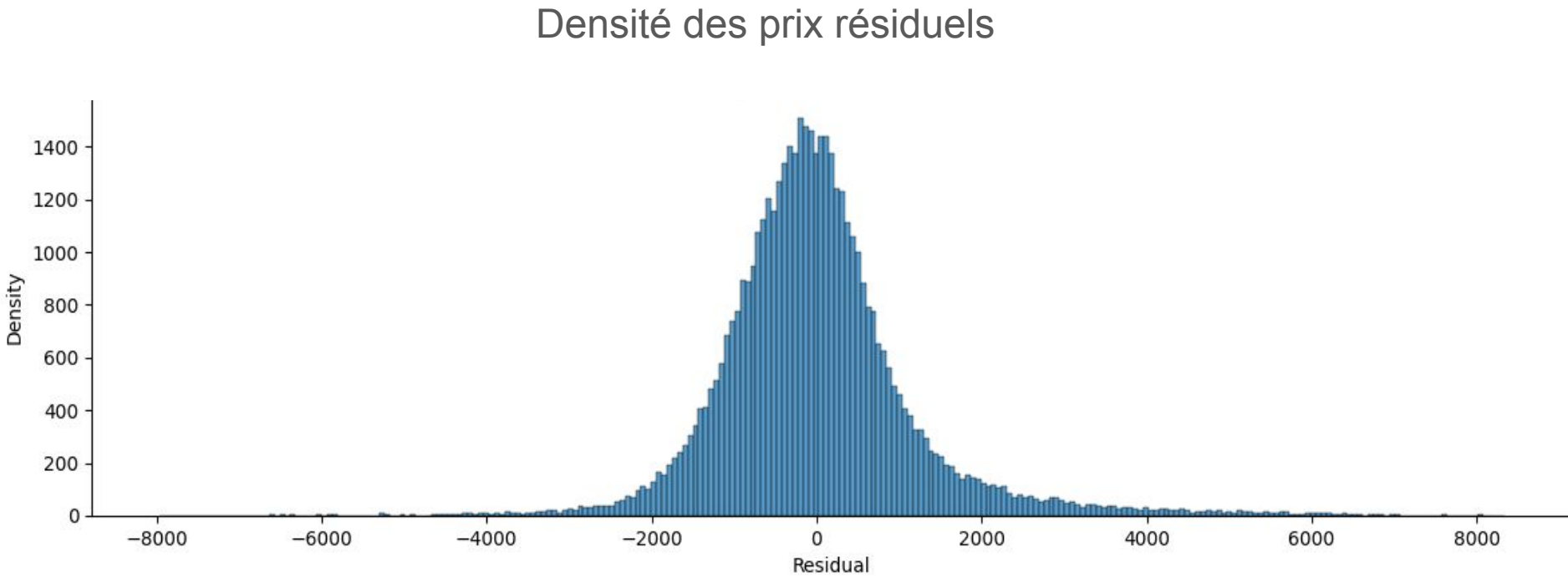
Polynomial Regression (séparation type et tranche)



Polynomial Regression (sans séparation)


```
Mean Absolute Error (MAE): 835.13
R-squared (R2): 0.56
```

Polynomial Regression (sans séparation)



Code

- Clean_data
- Mean_k_neighbors_price
- Split_train_test
- Polynomial_regression

Piste d’améliorations

- Prendre en compte la densité et élargir ou non le voisinage en conséquence
- Utiliser d’autres données
  - L’accessibilité aux transports de la commune
  - Le taux de chômage
  - Le taux de criminalité
  - Le nombre d’étages pour les appartements et la présence d'ascenseur
  - L’année de construction des biens
- Split train/test
  - Faire en sorte de représenter toutes les tranches de m carré dans les données train et test

Contributions

- **Idée du projet**: Prof
- **Beaucoup d'idées reprise des articles :**
  https://hureauxarnaud.medium.com/projet-estimateur-de-prix-dun-bien-immobilier-bas%C3%A9-sur-du-machine-learning-ae578fdacaca
  https://www.kaggle.com/code/auradee/house-price-prediction-using-polynomial-regression
- **Code**:
  - Partie "split_train_test": **Tharsiya et Marilyn**
  - Partie "random_forest": **Marilyn**
  - Partie "prix du voisinage": **Tharsiya**
  - Partie "polynomial regression" : **inspirée d'un projet Kaggle (lien ci-dessus), adaptée par Tharsiya**
- **Nettoyage des données**: Surtout **Marilyn**
- **Tuning** : **Tharsiya**
- **One hot encoding** : **Marilyn**
- **Algo BallTree (prix voisinage)** : **Tharsiya**
- **Graphes: Tharsiya**
- **Slides** (à part les graphes): **Surtout Marilyn**

Part de la contribution : 50% Tharsiya 50% Marilyn