



M2 SISE
Projet ML
Rapport d'analyse

Octobre 2023

Anthony Sardellitti
Nousra CHAIBATI - Abdourahmane NDIAYE - Fiona STEINER

Table des matières

1. Contexte.....	3
2. Traitement des données.....	3
a. Analyse - Exploration.....	3
b. Nettoyage.....	4
3. Enrichissement des données.....	6
4. Modèles prédictifs.....	6
a. Classification - Type local.....	6
b. Régression - Valeur foncière.....	7
5. Cartographie.....	8
a. Carte statique.....	9
b. Carte dynamique.....	9
6. Conclusion.....	10

1. Contexte

GreenTech Solutions est une société de service qui développe des applications. Une agence immobilière sollicite l'entreprise afin d'aider ses commerciaux à mieux estimer ces biens à vendre.

Cette application permet de mieux comprendre le marché et d'estimer le prix de vente des nouveaux appartements.

Les données fournies proviennent de la Direction générale des Finances publiques (DGFiP) qui rend accessibles au public, sur le site data.gouv.fr, les éléments d'information des valeurs foncières déclarées au cours des cinq dernières années.

Dans ce rapport, nous détaillons l'analyse qui nous a permis de prédire différents critères.

2. Traitement des données

Nous avons travaillé sur les années 2018 à 2021 pour notre analyse. La prédiction de la valeur foncière, quant à elle, a été effectuée sur l'année 2022.

Les fichiers fournis comportent au total 15 125 102 d'observations sur les quatre années pour 43 critères.

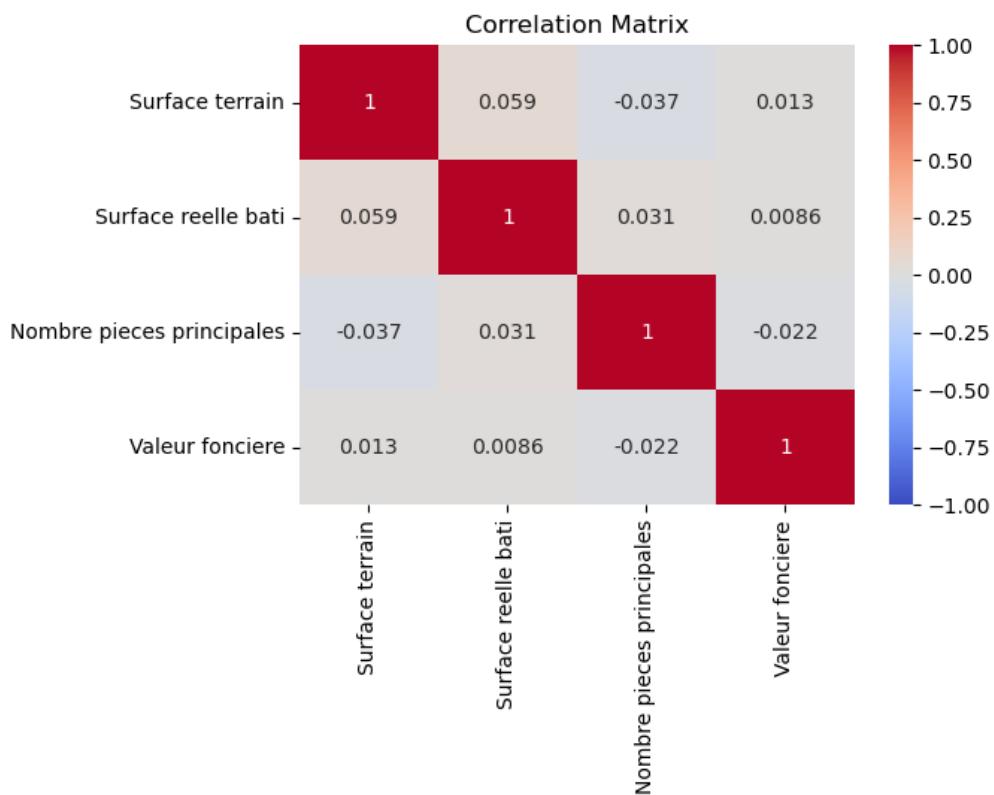
a. Analyse - Exploration

Avant de nous lancer dans la conception du modèle d'apprentissage, il était essentiel d'explorer les données afin de comprendre comment elles se présentent, d'identifier de potentielles relations et les éventuelles données aberrantes. Pour cela, on effectue une analyse d'exploration à l'aide de statistiques descriptives et de visualisation des données.

Plus concrètement, nous avons commencé sur un échantillon des données. D'abord, nous avons analysé le fichier des valeurs foncières de l'année 2018, ensuite, nous avons généralisé sur toutes les données. Parmi les analyses qui ont été réalisés, il y a :

- Affichage des informations globales sur le jeu de données :
 - Types des variables : 15 variables réelles, 4 variables entières et 24 variables en chaînes de caractères
 - Taille du jeu de données brut : 15 125 102 x 43
- Visualisation du pourcentage des données manquantes :
 - 21 variables à plus 90% vides
 - 8 variables entre 30 et 45 % de valeurs vides
 - 13 variables à moins de 2% vides
 - 1 variable à 69 % vides

- Matrice de corrélation les variables :
 - Surface terrain
 - Surface réelle bâtie
 - Nombre pièces principales
 - Valeur foncière



La matrice ci-dessus nous montre qu'il n'y a pas corrélation entre ces variables.

b. Nettoyage

Après avoir effectué l'analyse exploratoire, nous avons une vision claire des transformations que nous devons faire pour aboutir à un jeu de données propre et exploitable pour faire notre apprentissage.

Voici les transformations que nous avons appliquées :

- Les variables explicatives :
 - Supprimer des variables qui contiennent plus de 90% de valeurs manquantes
 - Supprimer les lignes pour lesquelles les colonnes suivantes sont des NA :

- No voie
 - section
 - Nombre pièces principales
 - Type local
- Remplacer les valeurs manquantes des variables suivantes :
 - Nature culture spéciale : par une nouvelle modalité appelée “NON DEFINI”
 - Nature culture : par une nouvelle modalité “NON DEFINI”
 - Type de voie : par une nouvelle modalité “NON DEFINI”
 - Surface terrain : par 0
 - Convertir des types des variables suivantes :
 - No voie : en type entier
 - Code type local : en type entier
 - Nombre pièces principales : en type entier
 - Code postal : en type chaîne de caractère
 - Supprimer les ventes de biens répétées plusieurs fois, effectuées à la même adresse et à la même date en se basant sur la concaténation des variables :
 - Voie
 - Code voie
 - Code postal
 - Date mutation
- La variable à prédire “Valeur foncière” :
 - Remplacer les virgules séparant la partie entière et la partie décimale par des points pour correspondre au type de données float du langage Python.
 - Convertir la colonne du type initial “object” au type “float”.
 - Supprimer toutes les lignes de ventes de biens vendus à l'euro symbolique ou gratuitement (Valeur foncière = 0 ou 1)

À l'issue de cette étape de nettoyage, le jeu de données nettoyé contient maintenant 2 315 484 observations et 22 variables et est prêt pour entraîner notre modèle d'apprentissage.

3. Enrichissement des données

Pour cette partie, nous avons enrichi le jeu de données avec des informations sur les ventes au mètre carré.

Ces données ont été récupérées sur le site du gouvernement qui propose un grand choix d'API.

Une colonne **PrixMoyen_M2** par département a été ajoutée pour affiner notre modèle.

4. Modèles prédictifs

a. Classification - Type local

Notre objectif final est de pouvoir prédire les valeurs foncières des biens vendus en 2022 à partir du modèle d'apprentissage que nous aurons obtenu après l'avoir entraîné sur les données des quatre années précédentes.

Par ailleurs, nous devons d'abord compléter les valeurs de type local manquantes des données de 2022 qui sont au nombre de 50 000.

Nous avons utilisé un modèle de classification KNN utilisant la méthode des K plus proches voisins pour prédire les valeurs de type local manquantes.

Ce modèle a entraîné sur 80% des données le jeu de données nettoyé et enrichi avec des données sur prix moyen au m2 par département, les 20% restants serviront d'échantillon de test.

Nous avons fait le choix des variables suivantes pour créer ce modèle :

- Prix moyen au m2
- Surface réelle bâti
- Nombre pièces principales
- Surface terrain
- Nombre de lots
- Type local

Pour la variable Type local n'avons retenu que les modalités "Appartement", "Maison" et "Dépendance" ensuite, nous l'avons binarisé pour l'apprentissage puisqu'il s'agit d'une variable qualitative.

Avec un nombre de voisins K fixé à 4, il est capable de prédire le type local avec un f1 score de 0, 97 sur l'échantillon de test.

b. Régression - Valeur foncière

Une fois notre jeu de données complété, nous avons fait le choix de quelques variables qui semblaient pertinentes pour prédire la valeur foncière d'un bien. Les variables qui ont été retenus pour notre modèle sont :

- Le prix moyen au m²
- La surface réelle bâti
- Le nombre de pièces principales
- La surface du terrain
- Le mois
- Le nombre de lots
- Le type de local

Nous avons choisi de binariser le type de local, car il s'agit d'une variable qualitative. Nous avons conservé trois de ses modalités qui sont : Appartement, Maison et Dépendance. Une troisième

Nous avons par la suite séparé notre jeu de données en données d'entraînement (80%) et données de test (20%).

Pour construire notre modèle, nous avons utilisé la méthode des arbres de régression et nous avons effectué l'apprentissage sur les données d'entraînement.

Nous avons testé des performances de notre modèle sur l'échantillon réservée à cette fin et nous avons eu les résultats suivants :

- Mean Absolute Error: 88340.90
- Mean Squared Error: 1872602875785.38
- RMSE: 1368430.8078179844
- R-squared: -0.01

Nous avons nos modèles, nous pouvons passer maintenant à la prédiction des valeurs foncières de 2022 :

1^{re} étape : Ajout et nettoyage des données 2022

Le nettoyage des données à prédire est une étape primordiale pour que les données respectent le format du modèle de prédiction :

- Conversions des types des variables
- Gestion des NA en les remplaçant par des valeurs ("NON DEFINI" pour les variables qualitatives et 0 pour les variables quantitatives)
- Enrichissement avec les prix moyens au m² par commune
- Binarisation du type local en conservant les modalités (Appartement, Maison, Dependance et NON DEFINI)

2^e étape : Prédictions les Type local manquants

Une fois les données de 2022 prêtées, on prédit le Type local.

3^e étape : Prédictions des valeurs foncières

Dans cette étape, on prédit les valeurs foncières des biens avec le jeu de données obtenu après la prédiction des 50 000 types locaux qui manquaient.

5. Cartographie

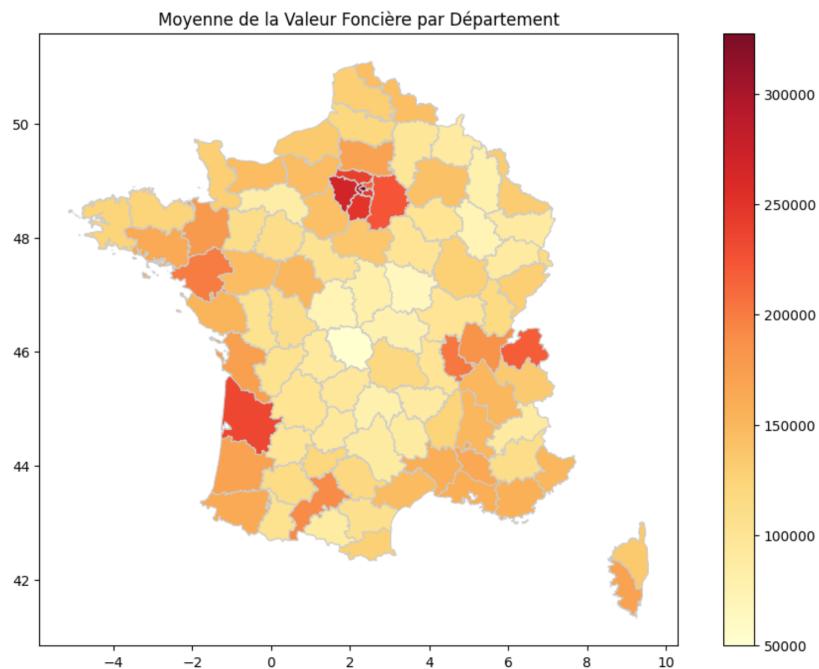
Afin d'avoir une représentation plus visuelle, une cartographie des valeurs foncières a été créée.

Pour ce faire, les données sur les contours des départements français ont été ajoutés au projet. Chaque département est représenté par un polygone.

Note : Les données des départements Haut-Rhin et Bas-Rhin ne nous ont pas été fournis.

a. Carte statique

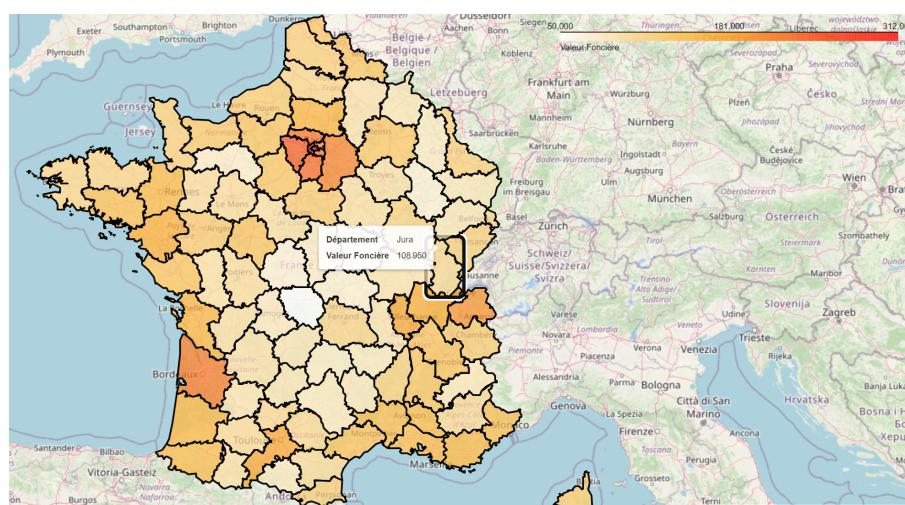
Une première carte, statique, comportant une légende a été créée.



b. Carte dynamique

La deuxième carte, dynamique, est superposée à la carte du monde et contient également une légende. Il est possible de naviguer sur la carte afin de visualiser les données par département.

Nous avons privilégié cette carte qui nous apporte plus d'informations.



6. Conclusion

Cette analyse nous a permis dans un premier temps de comprendre les données pour pouvoir les traiter. Dans un second temps, les nettoyer et prédire les **type local** manquant ainsi que les **valeurs foncières** de l'année 2022.