Real Estate Price Prediction in France

Machine Learning Approach to DVF Data

Willy Vo Matheo Quatreboeufs Donovan Thaing

Master 2 – IASD

Introduction

Main question

How can we predict the price of a real-estate property in France based on its characteristics and its geographical environment ?

Key Challenges :

- A complex and heterogeneous real-estate market
- Strong influence of geographical factors (proximity to services, transportation, etc.)
- Public data is available but scattered across multiple sources

Project Objectives

Main goal

To build a complete data-processing and enrichment pipeline for French real-estate information, from downloading raw data to generating a final dataset ready for predictive real-estate price modelling using machine learning. To integrate those data in databases.
To collect and clean property transaction records (DVF), geographic points of interest (OpenStreetMap), macro-economic indicators (Banque de France), and socio-economic statistics (INSEE).

1. To **Collect** and **clean** property transaction records (DVF)
2. To **Enrich** with geographic points of interest (OpenStreetMap), macro-economic indicators (Banque de France), and socio-economic statistics (INSEE).
3. **Build** an efficient predictive model and **analyse** the decisive factors
4. To **construct** databases to store those data

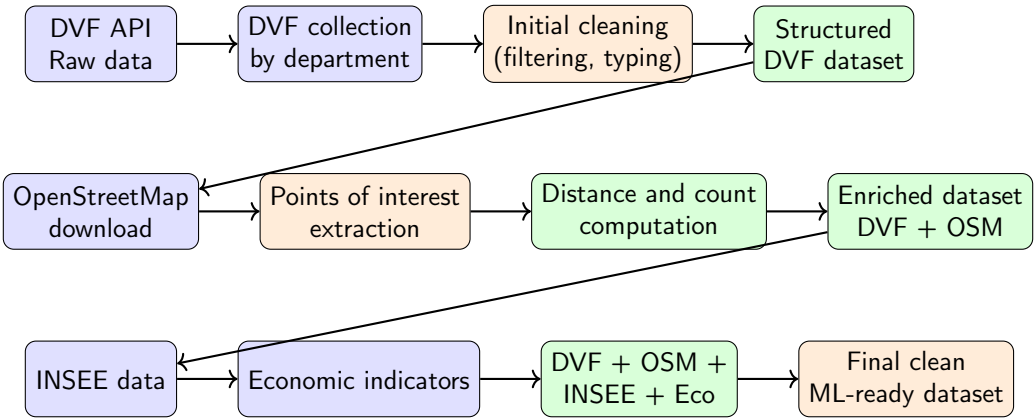
Data sources overview

Our dataset is built by combining several official and open data sources, each providing complementary information about the real estate market.

- **DVF (data.gouv.fr) :** detailed real estate transactions. (val. fonciere, maison, etc...)
- **OpenStreetMap :** geographical and environmental context (shops, schools nearby etc.).
- **INSEE :** socio-economic indicators at local scale. (commune / department level etc.)
- **Macroeco. data (Banque de France) :** national economic conditions over time (inflation etc.).

We collected the data through REST APIs and by downloading CSV files directly from the official websites.

Data Pipeline



Data cleaning : main phases

Initial cleaning (DVF transactions)

- Filter transactions to keep only actual property sales.
- Standardize prices, surfaces, and dates into consistent formats.
- Build basic and reliable property-level indicators.

Spatial cleaning (OpenStreetMap enrichment)

- Select relevant categories of points of interest and spatial scales.
- Compute the numbers of POI in a radius and the nearest POI. (POI = station, schools, shops ...)

Final cleaning (merged dataset)

- Harmonize variables from heterogeneous data sources.
- Handle remaining missing values and extreme observations.
- Apply final transformations to obtain an ML-ready dataset.

OpenStreetMap : extracting spatial features from POIs

Principle :

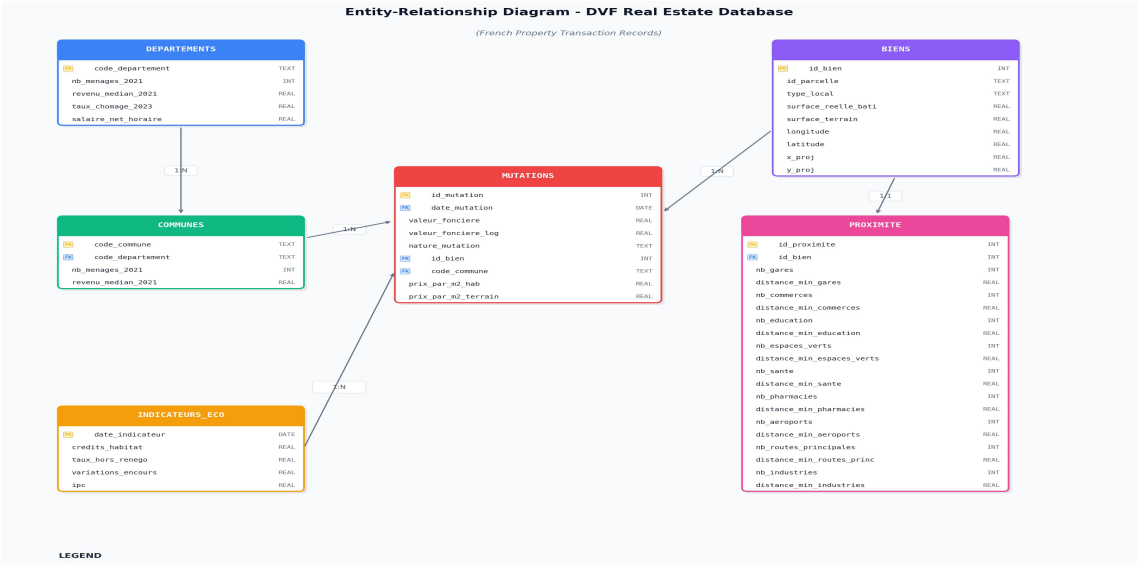
- Identify relevant **Points of Interest (POIs)** around each property.
- For each POI category :
  - count the number of POIs within a fixed radius,
  - compute the distance to the nearest POI.

Extracted POIs and spatial parameters :

POI category	Radius (m)	OSM tags
Train stations	1500	`public_transport=station`
Shops	500	`shop=supermarket|convenience`
Education	1000	`amenity=school|university`
Green areas	500	`leisure=park`
Health facilities	1000	`amenity=hospital|clinic`
Airports	30000	`aeroway=aerodrome`

These features capture both proximity and density effects of amenities, which are known to

Database DVF



Database DVF

DB Browser for SQLite - C:\Users\donov\PycharmProjects\DataExtraction\deferla.db

Fichier Édition Vue Outils Aide

Nouvelle base de données | Ouvrir une base de données | Enregistrer les modifications | Annuler les modifications | Annuler | Ouvrir un projet | Enre

Structure de la base de données | Parcourir les données | Éditer les pragmas | Exécuter le SQL

Table : annonces — Filtrer dans n'importe quelle colonne

	id	url	date_publication	type	titre
	Filtre	Filtre	Filtre	Filtre	Filtre
1	86382073	https://www.deferla.com/bien?…	2025-10-20	Appartement	Studettes à vendre
2	85957542	https://www.deferla.com/bien?…	2025-05-02	Appartement	Paris XIVème- Rue Louis Morard - 2 …
3	86121998	https://www.deferla.com/bien?…	2025-05-14	Appartement	Paris XIVème - Rue Louis Morard - …
4	85820745	https://www.deferla.com/bien?…	2024-12-09	Appartement	Appartement familial 319 m² - Neuill
5	86122395	https://www.deferla.com/bien?…	2025-06-25	Parking	Italie II - emplacement de parking
6	85957238	https://www.deferla.com/bien?…	2025-03-31	Parking	rue Daguerre - emplacement de parkir
7	86122111	https://www.deferla.com/bien?…	2025-05-24	Parking	Montparnasse - double emplacement de
8	86388970	https://www.deferla.com/bien?…	2025-10-23	Parking	Paris XIVème: St Jacques - Dareau: U
9	86353594	https://www.deferla.com/bien?…	2025-10-07	Appartement	Grand studio avec vue dégagée - …
10	85957293	https://www.deferla.com/bien?…	2025-04-05	Appartement	Villa de Lourcine - Appartement …
11	86462032	https://www.deferla.com/bien?…	2025-11-20	Appartement	Appartement cosy à Pernety - Paris …
12	85956938	https://www.deferla.com/bien?…	2025-02-26	Parking	Parking - Cardinal Lemoine
13	85819781	https://www.deferla.com/bien?…	2024-09-24	Appartement	Appartement 4 pièces - XIVème PARIS
14	85819779	https://www.deferla.com/bien?…	2024-09-24	Appartement	Appartement 4 pièces - XIVème PARIS
15	85819782	https://www.deferla.com/bien?…	2024-09-24	Appartement	Appartement 4 pièces - XIVème PARIS
16	85819777	https://www.deferla.com/bien?…	2024-09-24	Appartement	Appartement 4 pièces - XIVème PARIS
17	85819776	https://www.deferla.com/bien?…	2024-09-24	Appartement	Appartement 4 pièces - XIVème PARIS
18	85819775	https://www.deferla.com/bien?…	2024-09-24	Appartement	Appartement 3 pièces - XIVème
19	86339783	https://www.deferla.com/bien?…	2025-09-30	Parking	Place de parking - Paris VIème
20	86122588	https://www.deferla.com/bien?…	2025-07-17	Appartement	Studette - XVème PARIS

Context The Technical Problem

Target Selection

- **Goal :** Scrape real estate data.
- **Constraint :** Major sites (SeLoger, LeBonCoin) block scraping via `robots.txt`.
- **Decision :** We targeted a smaller agency, **"De Ferla"**, to respect strict web rules.

The JavaScript Roadblock

- **Issue :** The website uses JavaScript to display listings.
- **Result :** Standard HTML scraping (XPath) did not work.

The Solution : API Reverse Engineering

Methodology

To fix the JavaScript issue, we analyzed the **Network Traffic**.

- We searched for the hidden link used by the site.
- **Discovery :** We found the direct API endpoint ("The Holy Grail").

The Result

- **Efficiency :** We obtained **all data** with just **one request**.
- **Quality :** The data was very clean and structured (JSON).
- **Limitation :** We retrieved image *links*, but not the image *files* yet.

Automation : Scrapy Image Pipeline

We used **Scrapy** to automate the process and handle files.

Custom Image Pipeline

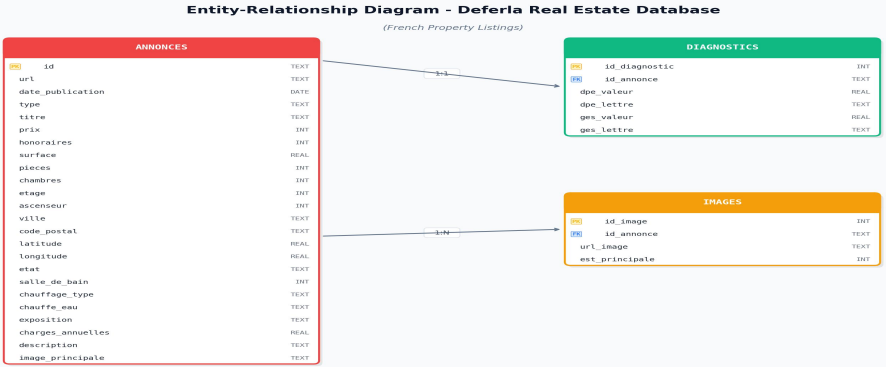
We defined a specific logic for the images :

- **Download :** Convert links into files automatically.
- **Organization :** Rename every image using property ID.
- **Storage :** One folder per property for better structure.

Privacy GDPR

- We deliberately excluded personal details.
- **No Phone/Email :** We did not scrape agent contact info to respect privacy rules.

Database Deferla



Database Deferla

DB Browser for SQLite - C:\Users\donov\PycharmProjects\DataExtraction\dvf_immobilier.db

Fichier Édition Vue Outils Aide

Nouvelle base de données | Ouvrir une base de données | Enregistrer les modifications | Annuler les modifications | Annuler | Ouvrir un projet | Enreg

Structure de la base de données | Parcourir les données | Éditer les pragmas | Exécuter le SQL

Table : biens | Filtrer dans n'importe quelle colonne

	id_bien	id_parcelle	type_local	surface_reelle_bati	surface_terrain	longitude	latitude	x_proj	y_proj
	Filtre	Filtre	Filtre	Filtre	Filtre	Filtre	Filtre	Filtre	Filtre
1	1	01422000ZK0154	Maison	109.0	1057.0	5.315595	46.134815	878746.066218434	6562066.2196
2	2	01322000AS0816	Maison	82.0	467.0	4.801234	45.935239	839554.759414373	6538873.9046
3	3	011850000F0233	Maison	105.0	436.0	5.589117	45.974042	900437.998194435	6544877.6068
4	4	010990000D1319	Maison	117.0	1005.0	5.263891	45.883244	875556.881992233	6534024.8849
5	5	013140000E1869	Maison	98.0	212.0	5.28346	46.001436	876694.461468944	6547188.353
6	6	01249000AC0559	Maison	188.0	715.0	4.94681	45.826186	851129.945509211	6527034.8877
7	7	010080000A2489	Maison	110.0	610.0	5.334509	45.939925	880845.424264444	6540476.030
8	8	01082000AN0272	Appartement	54.0	0.0	5.781224	46.003311	915189.675379245	6548631.6513
9	9	012460000B0422	Maison	97.0	959.0	5.044841	46.238913	857551.593941847	6573045.4898
10	10	010740000E1037	Maison	124.0	677.0	5.166759	45.99903	867673.779901403	6546666.7959
11	11	01249000AE0489	Appartement	61.0	0.0	4.961159	45.825556	852245.360860407	6526992.511
12	12	01053000AZ0061	Maison	220.0	245.0	5.23075	46.200397	871992.7407084	6569157.7286
13	13	01053000BN0267	Maison	109.0	600.0	5.231504	46.216907	871999.052001091	6570992.1000
14	14	01160000AD0097	Appartement	66.0	0.0	6.102157	46.251811	938925.004562657	6577131.5805
15	15	012900000B0796	Appartement	47.0	0.0	5.18603	45.900261	869468.292253454	6535743.7725
16	16	013780000D1110	Maison	49.0	126.0	5.185712	45.810816	869718.557746984	6525813.9894
17	17	01032000C3103	Maison	106.0	729.0	5.128262	45.856001	865123.484144998	6530708.1251
18	18	01416000AL0076	Maison	150.0	511.0	5.511072	45.937168	894528.365945743	6540589.683
19	19	012520000B0459	Maison	132.0	500.0	4.810537	46.139918	839754.279006642	6561614.1263
20	20	013900000D1077	Maison	35.0	5700.0	5.284495	45.822754	877348.524680521	6527356.2726

Questions ?

Thanks for your attention !