Immo Eliza: Regression Challenge

Group: Tequila

Contributors:

- Sofia
- Marc
- Moussa
- Kenny

Preprocessing Overview — Key Steps

Our preprocessing workflow involved several critical steps to prepare the data for modeling:

Dropped Irrelevant or Duplicate Features

Removed: ID, URL, and Municipality (duplicate of postal code)

Removed Highly Correlated Variable

Excluded: Price per m² due to strong correlation with total price

Categorical to Numerical Conversion

EPC Score: A+ to G → mapped to **1–9**

Building Condition: Categories → mapped to **1–5** scale

Data Type Adjustment

Converted Postal Code to string (for target encoding)

Train-Test Dataset Split

Training Set: 80%

Test Set: 20%

Fixed split reused across models for fair comparison

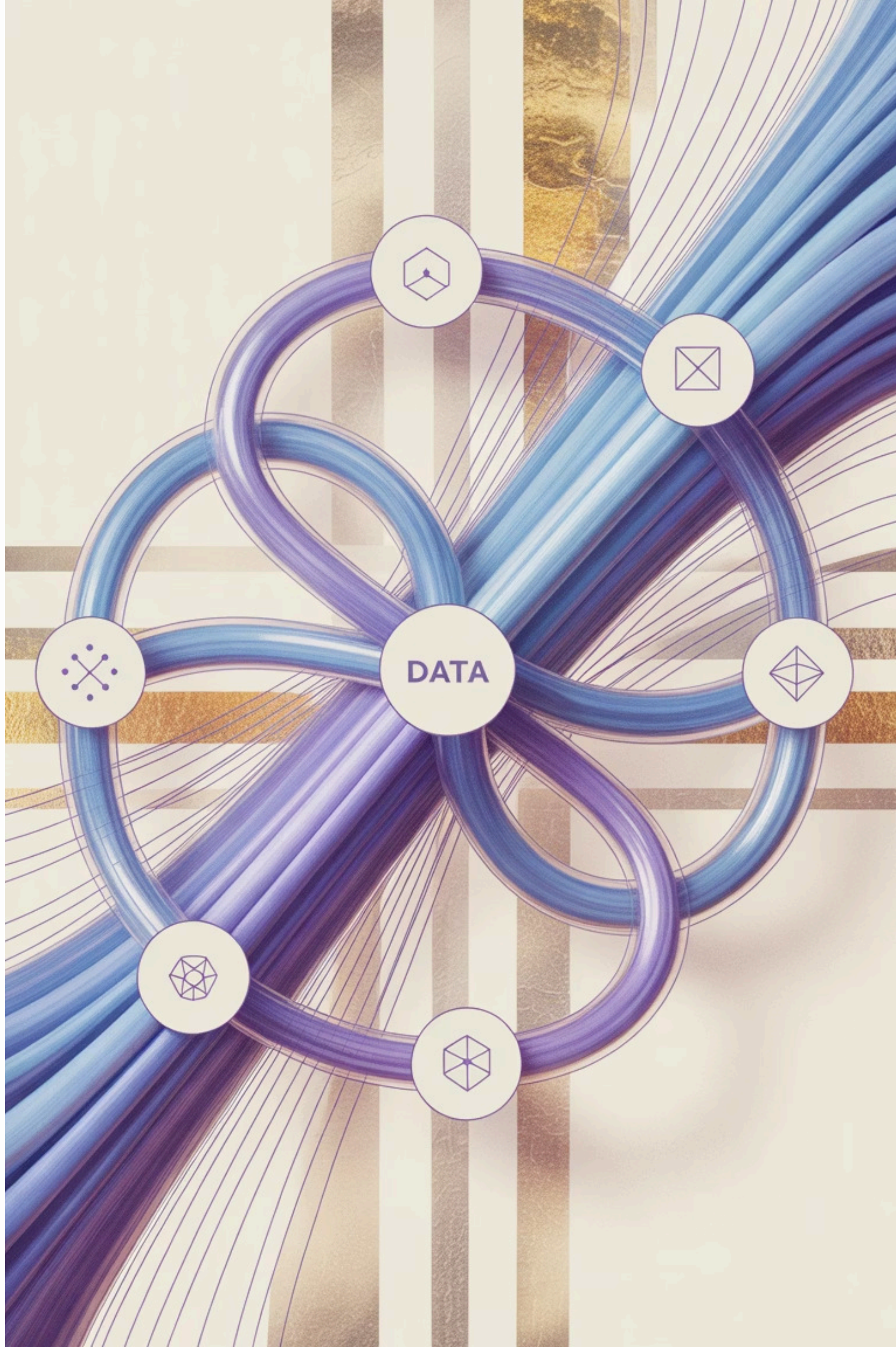
Pre-Cleaned Dataset with Outlier Capping

Price: €50,000 – €1,000,000

Habitable Surface: 15–600 m²

Bedrooms: Max 20

Garden Surface: Max 3,000 m²



Investigation — Model Exploration

Models Tested

9 models explored: Multiple Linear Regression, Random Forest, XGBoost, Lasso, GradientBoost, CatBoost, LightGBM, Ridge, Elasticnet

Explored multiple **variants** of each model

Built and evaluated **stacked models**

Hyperparameter Tuning

Used **Optuna** to optimize parameters, particularly for CatBoost

Feature Analysis

Benchmarked models using **all features vs top-ranked features**

Segmentation Strategies

Created targeted models by **segmenting data** by:

- Property Type
- Region
- Subtype

Train/Test Evaluation

All models evaluated on **both training and test sets**

Ensured **no overfitting** through consistent testing

CatBoost Model Highlights

Gradient Boosted Trees

Learns by correcting residuals step-by-step

Handles categoricals natively

No encoding needed

Ordered Boosting

Avoids overfitting on small categories

Feature importance

Includes interactions for deeper insights

Tuned with Optuna

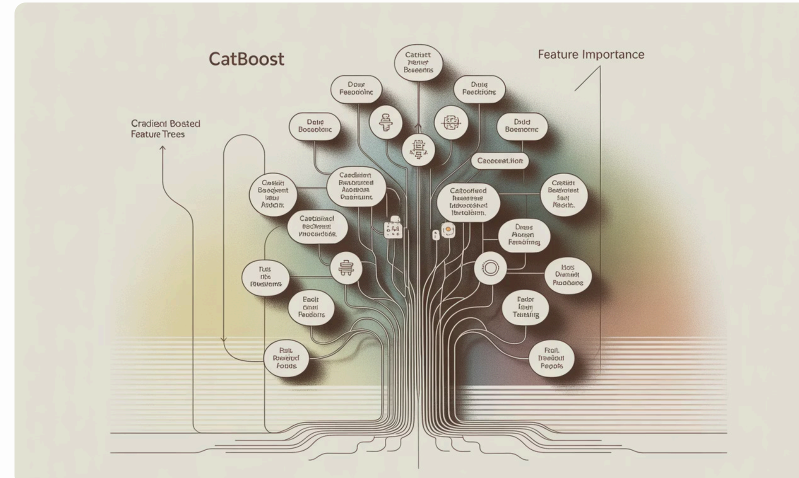
Best RMSE using custom loss

Minimized Loss Function

Tested MAE, RMSE inside CatBoost

K-Fold CV used

For stable evaluation



Possible next steps

- Try SMOTE for rare subtypes
- Add GridSearchCV for benchmark comparison
- Test KMeans clustering for regional segmentation
- Add ROC-AUC/MCC for binary/value band classification

Offers clearer interpretability than traditional approaches (e.g., XGBoost)

Results — Best Scores

€58,355

Best Overall MAE

Lightgbm + Optuna

€84,986

Best Overall RMSE

Lightgbm + Optuna

0.797

Best Overall R^2

Lightgbm + Optuna

Best Segment Performance (Apartment)

Model	Metric	Value
CatBoost	MAE	€49,283
CatBoost	RMSE	€76,524
CatBoost	R^2	0.784

Stacked Model Performance

Ensemble of **XGBoost + Ridge + GradientBoost**

Achieved R^2 = 77.3%

Challenges & Reflections

Key Challenges

Imbalanced data

For specific subtypes of properties, such as Castle and Penthouse

Feature Leakage Risk

Price per m^2 was too closely tied to target variable -> removed ppm^2

Overfitting

In certain models, such as random forest, where the train r2 was significantly better than the test r2.

SHAP Analysis

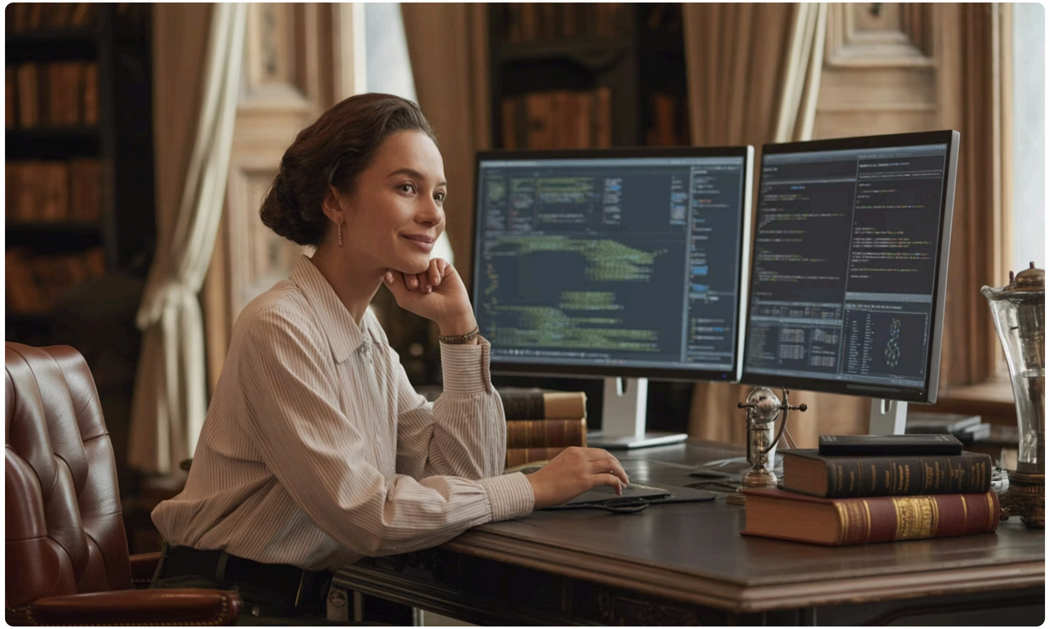
Provided valuable explainability, revealed feature dependencies

Optuna Tuning

Improved performance, but computationally expensive

Segmentation Trade-Offs

Increased model accuracy at the cost of longer training times



Next Steps & To-Do

- **RFI (Recursive Feature Inclusion)** to further refine feature selection
- Explore loss function optimization techniques to **minimize errors**
- Try **clustering with z-score normalization** for latent pattern discovery