Hellenic Republic
Ministry of Finance

Analyzing Greek real estate transaction data with ML techniques

Preliminary results – Comments are welcome

Veni Arakelian, Council of Economic Advisors, MinFin,

&

Myrto Gasparinatou, Ionian University & Fujitsu Luxembourg

Introduction

- The housing market plays a critical role in the economy, influencing GDP, financial stability, and household wealth.
- This study uses ML algorithms to identify the key determinants of residential property prices in Greece.

Data Overview

- Source: *Greek Real Estate Transactions Valuation Registry* (2017–2024)
- Initial entries: 255,368
- We restrict the dataset to residences or apartments and detached houses, obtaining 139,235 entries, excluding e.g. land, athletic facilities, etc.

Data Overview

- For each property, the following characteristics are recorded:

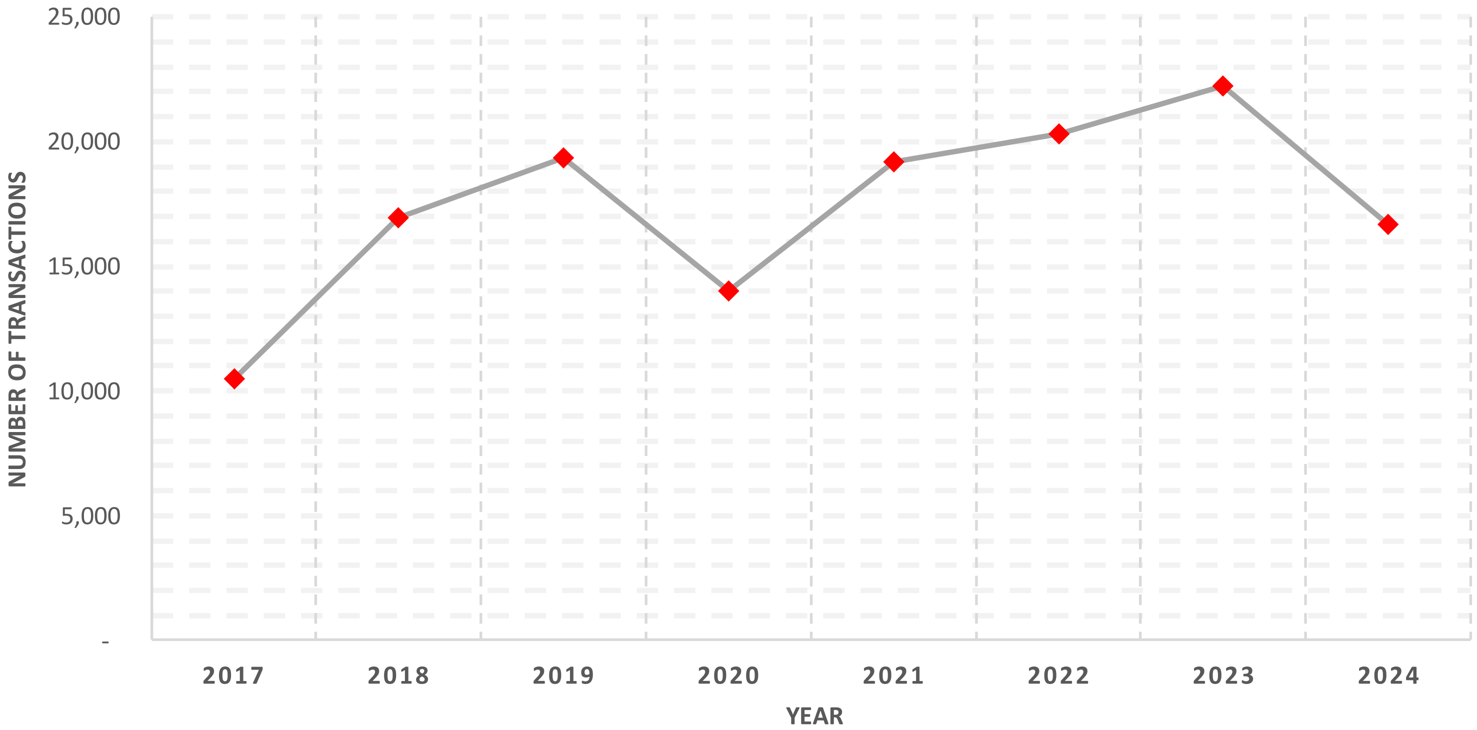
  prefecture, municipality, ***type***, ***number of facades***, ***zone price***, ***main area surfaces (in sqm)***, ***year of construction***, ***ownership percentage***, ***floor***, ***transaction date***, and ***price sold***.

- Cleaned dataset excluding nulls: 75,972 residential transactions

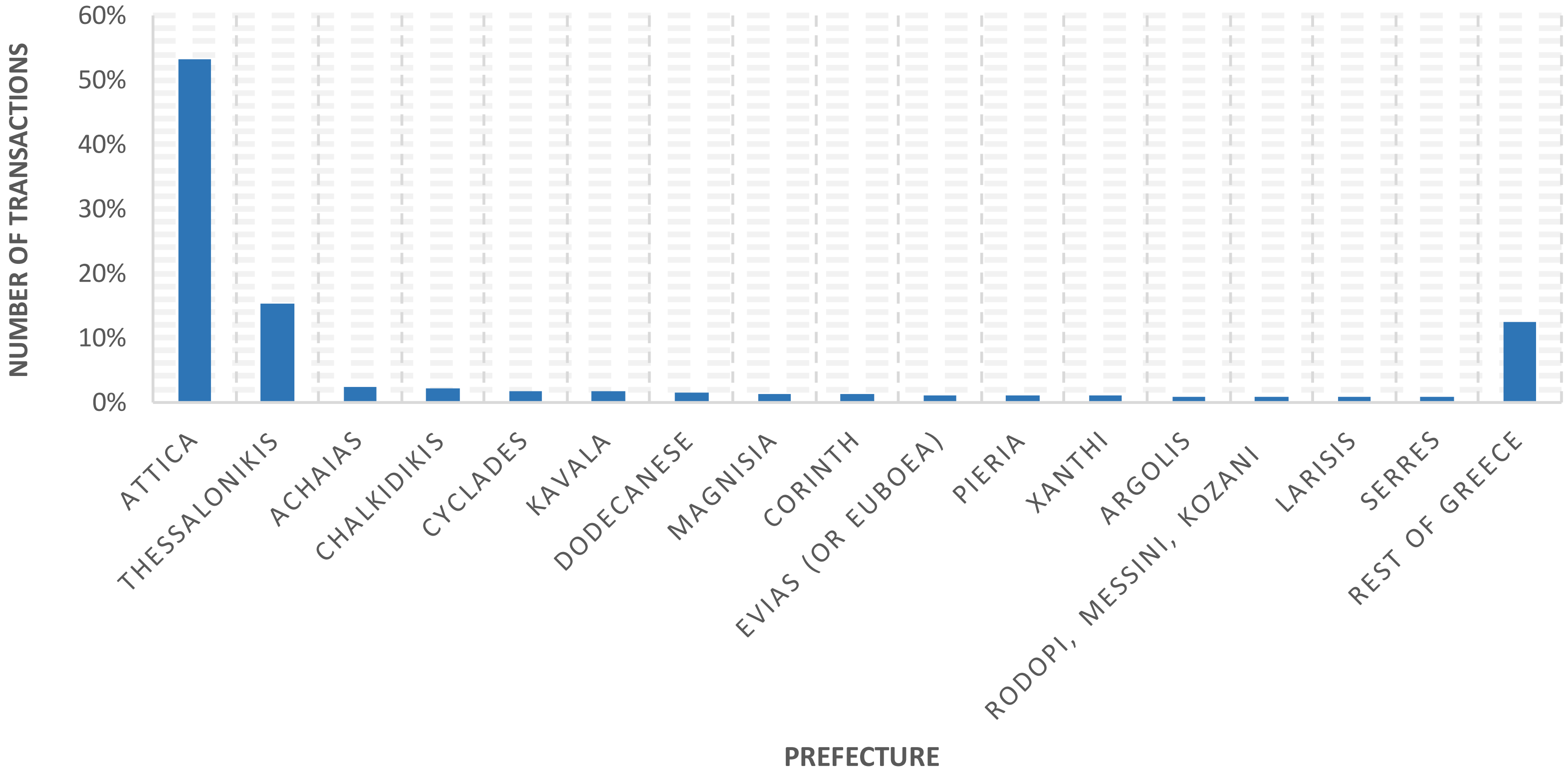
- Features: 918 (quantitative + location-based dummies)

- Target feature: Price per sqm (log-transformed) calculated as the ratio of the price sold over the main area surfaces.

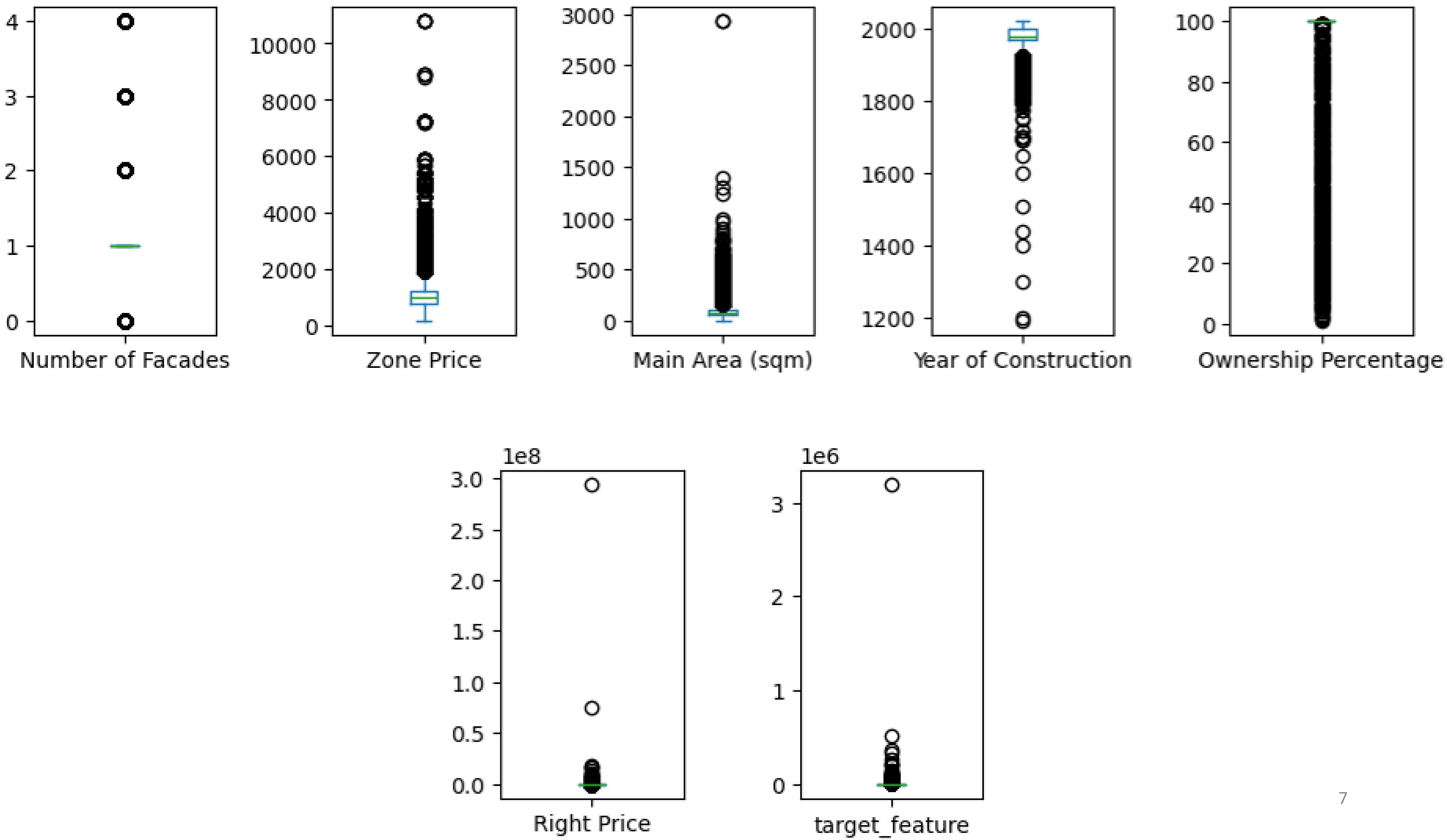
Data Overview-Number of transactions per Year



Data Overview- Number of transactions per Prefecture



Data Overview



Modeling Framework

- Models: OLS, Random Forest, Gradient Boosting, XGBoost
- Evaluation: R^2 and Mean Squared Error (MSE)
- Feature selection: mRMR Feature Selection Method and feature importance-based selection *(work in progress)*

Key Modeling Equations

- OLS Regression: $y_i = \beta_0 + x_i' \boldsymbol{\beta} + \varepsilon_i$
- Random Forest Prediction: $\hat{y}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x)$
- Gradient Boosting: $F_M(\boldsymbol{x}) = \sum_{m=0}^{M} \eta h_m(\boldsymbol{x})$, where is the regression tree fitted to the pseudoresiduals
- XGBoost Leaf Weight: $w_j^* = \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$, where g_i, h_i are the first and the second order gradients of the loss fuction, and I_j indexes samples in leaf j.

Results: Full Feature Set

Model	Mean R^2	Mean MSE
Linear Regression	**0.586**	**17,267,150**
Gradient Boosting	0.572	119,423,300
Random Forest	0.477	119,760,000
XGBoost	0.186	104,243,465

Feature Importance: Top 10

Model	Mean R^2	Mean MSE
Linear Regression	**0.590**	**17,190,420**
Gradient Boosting	0.569	117,409,700
Random Forest	0.498	118,382,000
XGBoost	0.078	104,288,238

Municipalities as Top 10 Features

Khanion
Messarion
Vouliagmenes
Agiou Ioannou Rente
Oias
Voulas
Patmou
Theras (Santorini)
Messarion
Antiparou

Conclusion

- OLS outperforms ensemble models (RF, GB)
- Feature selection improves accuracy
- Linear models benefit most from reduced, relevant inputs

THANK YOU FOR ATTENDING!

mRMR Feature Selection Method

Minimum Redundancy Maximum Relevance, mRMR, seeks to maximize the difference betweern ***relevance*** and ***redundancy***, i.e.,

Average mutual information between feature pairs

$$mRMR(f_i) = I(f_i; c) - \frac{1}{|S|} \sum_{f_i \in S} I(f_i; f_j)$$

Mutual information between the feature and the class label

Main strengths:

- Balance between relevance and redundancy
- Mitigates the risk of selecting multiple features that convey overlapping information