THE DETERMINANTS OF NON-RESIDENTIAL REAL ESTATE PRICES:

A MACHINE LEARNING APPROACH

Raffaella Barone[♣]

Abstract

The aim of the paper is to establish a connection between the price of non-residential properties and indicators regarding social, economic and environmental characteristics of the province in which they are located. In particular, we wonder which of the indicators proposed by the Eni Enrico Mattei Foundation, and the Italian network of the Sustainable Development Solutions Network (SDSN Italia) to monitor the 17 sustainable development objectives in the various Italian cities, have the greatest impact in determining the price of non-residential properties built in the various Italian provinces. Moreover, as input variables we also considered the crime rate of the Italian provincial capitals, the per capita GDP and the frequency of sales of non-residential properties considered for the purposes of the analysis. After an analysis showing the landscape of the Italian non-residential real estate market, we used this dataset to train a supervised machine learning algorithm that was able to predict sale prices efficiently and accurately. We then used the trained model and the SHapley Additive exPlanations approach to assess the importance of each variable in the dataset with respect to the sale prices. We found that categorical variables and some SGD indicators are deeply linked to the outcome prices. Finally, to delineate a clear picture for possible policy measures, we used causal inference to understand how such interventions may affect prices.

Keywords: Machine Learning, Real estate market, Financial Stability, Sustainability, Crimes

JEL: B4: Economic Methodology, C1: Econometric and Statistical Methods and Methodology: General, G01: Financial crises, R33: Non-agricultural and Non-residential Real Estate Markets, K: Law and economics

---

[♣] Raffaella Barone, Department of Law, University of Salento; Baffi Centre, Bocconi University. E-mail: raffaella.barone@unisalento.it.

Acknowledgements: The author wishes to thank participants in The 5th International Conference on Corporate Governance and Risk Management – GRM 11-12 July 2024, which was organized by The University of Bari Aldo Moro and University of Foggia.

Funding: The work was supported by the Apulia Region and ARTI REFIN project cod. 00CA919A through the action 10.4 POR Puglia FESR FSE 2014–2020. The final responsibility is up to the author, without any involvement of the founding source.

1. Introduction

The residential and non-residential real estate sector constitutes a large share of the world's GDP: in US, between 2007 and 2020, it was equivalent to a median value of 15.2% of GDP at annual constant price; in UK it was equivalent to a median value of 17.5%; in European Union it amounted to a median value of 14%, with a 16.65% in the case of Italy.

In the last years, the analyses of the determinants of residential real estate price have generated an impressive strand of literature (Batalha et al. 2022; Beland et al. 2020; Gupta et al. 2022; Liu and Su 2021; Zhao 2020; Agarwal et al. 2020; Capponi and Rios 2020; D'Lima et al. 2021; Dave et al. 2020; Duca et al. 2021; Granja et al. 2020; Liu and Tang 2021). Few contributions used the agent-based model to study the determinants of house prices and the impact of money laundering on the amount and the prices of transactions in residential real estate market (Barone, 2023; Cincotti, 2022; Carro et al., 2022; Ozel et al., 2019; Baptista et al., 2016; Geanakoplos et al., 2012).

On the other hand, the contributions concerning the determinants of non-residential property prices are very limited. Most of them deserves particular attention to sustainability with a focus on green building constructions (Mangialardo et al., 2019; Bowman and Wills, 2008; Fuerst and Mcallister, 2008a, b). A small number of contributions have focused on the impact of environmental amenities on non-residential building constructions such as office, commercial, industrial, and retail property sales (see Franco and Cutter, 2022). However good environmental performance increases the property value (Porter and van der Linde, 1995; Klassen and McLaughlin, 1996). A change in environmental variables produces a variation on real estate prices and transactions.

In light of these observations, using a supervised machine learning methodology and following AGENDA 2030 and AGENDA 21, in this paper, we analyse the impact of the 17 sustainable development goals (SDGs) on the Italian non-residential real estate properties prices. Moreover, we enrich these databases also considering for each province the crime rate, the GDP and the frequency of sales, given by the ratio between the volume of the sales and the real estate stock.

Once the predictive model was created, and having assessed its performances, we continued the analysis with the evaluation of the features importance using the SHapley Additive exPlanation approach. Then, using DoWhy and EconML libraries for causal inference, we evaluated the cause-and-effect relation in data. To the purpose, we hypothesized that only a restricted number of the features have a causal relation with the outcome.

2. The Methodology

The implementation of the 2030 Agenda for sustainable development passes through the determination of the efficient level of exploitation of the environment. This assessment requires a cost-benefit analysis because environmental protection is not a free lunch. It would therefore be necessary to see how much individuals are willing to pay for environmental protection, i.e. what the demand for the environmental good is. The problem is that environmental goods are public goods. The market does not provide us with information on the quantity of the environmental good that is exchanged at different prices. Therefore, to determine the demand for the environmental good, scholars resort to direct and indirect methods of evaluation (*stated preferences* versus *revealed preferences*). One of the common indirect methods of evaluation is the hedonic price model (Griliches, 1961; Lancaster, 1966; Rosen, 1974; Lucas, 1975) which uses linear regression techniques. Applications in the real estate market have concerned the residential sector (see Hill, 2013 for a review of the literature), while the non-residential sector has been neglected.

In this contribution for property pricing, we use a machine learning algorithm, i.e. gradient boosting trees (for an analysis on the way in which machines learn, see De Liso, 2023) capable of efficiently capturing any non-linear dependencies. Potrawa, and Tetereva (2022) tried to demonstrate the superiority of the machine learning approach over the traditional hedonic regression model for the residential sector. However, machine learning models are often considered a black box due to the difficulty in interpreting the results. This challenge has recently been addressed using XAI (eXplainable Artificial Intelligence) methods. Chen et al. (2020) used the SHAP (see Lundberg S.M. and Lee S.I., 2017) for the analysis of the residential sector in Shanghai. However, the SHAP has a limit, namely that it can't be used to make transparent the correlations between variables, so it can't be used to infer any causal relation (De Villa, 2023). To overcome this limit, using the DoWhy and EconML libraries from the PyWhy project (Sharma and Kicirman, 2020), in this paper we performed the causal inference analysis merging two methods: 1) a type of graphs, called Directed Acyclic Graphs (DAG), which provide a visual representation of causal assumptions (Pearl, 1995; Pearl et al. 2016); 2) The Potential Outcomes (PO) framework (Splawa-Neyman, Dabrowska, and Speed, 1990; Rubin, 2005). The resulting causal estimate of potential outcome is usually called the average treatment effect (ATE).

3. Data collection and processing

The first step consisted of collecting, processing, and analysing the data.

- ***Non-residential real estate prices***

Firstly, we collected the prices relating to the following non-residential real estate properties: offices, shops, warehouses, laboratories, shopping centres, industrial warehouses, pensions and similar. Data were provided by the Real Estate Market Observatory (OMI)[1] of the Revenue Agency for the years 2016-2020. The prices are divided by provincial capital, urban area (B=centre, C=semi-centre, D=outskirts, E=suburban area, R=extra-urban area/agricultural area), year and conservation status of the property (excellent, normal, poor). For each property, province, conservation status and micro zone, the sales prices included a minimum price and a maximum price, of which we calculated the average value. Therefore, we defined the variable as "average purchase price".

Just as an example let's look at the case of the shops in Alessandria, a province in Piedmont (condition: normal). In semi-centre zone 'C' it is possible to identify three micro zones: C1, C2, C3. The minimum price of zone C for shops fluctuates between 1000 (€/sm) and 1300 (€/sm); while the maximum price of zone C fluctuates between 1300 (€/sm) and 1950 (€/sm). We have therefore considered a range of 1000-1300 (€/sm) as the minimum price of the zone and a range of 1300-1950(€/sm) as the maximum price of the area.

ALESSANDRIA C	C1	AL00000003	5	Shops	NORMAL	P	1050	1300
ALESSANDRIA C	C2	AL00004457	5	Shops	NORMAL	P	1300	1950
ALESSANDRIA C	C3	AL00004458	5	Shops	NORMAL	P	1000	1300

Then, for each zone (centre 'B', semi-centre 'C', outskirts 'D', suburban area 'E', extra-urban area/agricultural area 'R') we calculated the average purchase price.

- ***Real estate stock and volume of sales***

Secondly, we collected data relating to the real estate stock, provided by the revenue agency and available on the website:

https://www.agenziaentrate.gov.it/portale/web/guest/schede/fabbricatiterreni/omi/pubblicazioni/statistiche-catastali.

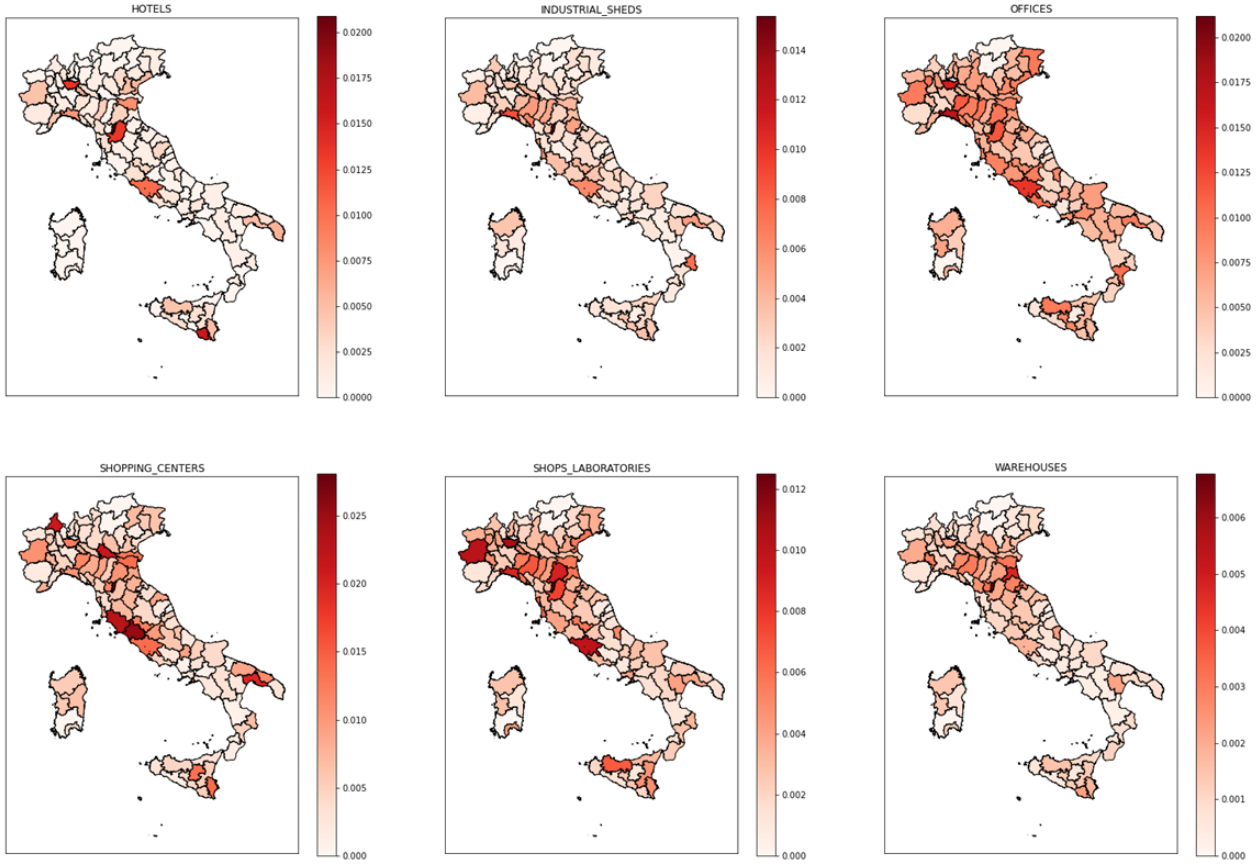
Furthermore, again from the revenue agency website, for each non-residential property that we are considering for each provincial capital, we collected the data relating to the volume of sales. Then we calculated the sales-to-stock ratio for each province, determining the frequency of sales. We

[1] Agenzia delle Entrate-Osservatorio del Mercato Immobiliare

plotted this distribution of sales frequencies using the geopandas library (in Python) and we obtained the graph shown in Figure 1

Looking at the data, it results that the most active cities were Prato (for all the non-residential real estate properties), Genoa (for offices, shops and laboratories and industrial sheds), Milan (for offices, shops and laboratories and hotels), Florence (for offices and hotels), Rome (for offices, shops and laboratories, industrial sheds, and hotels).

Figure 1: The average frequencies for each non-residential property



Source: Own elaboration based on data analysis

- ***Crime rate and GDP per-capita***

Moreover, as the level of crime in a city also impacts the value of properties, especially non-residential ones which are often the object of interest by organized crime for money laundering purposes, we took, from the Istat website, the data relating the crime rate of the provincial capitals and to GDP per-capita for the period 2016-2020. As the crime rate, we considered both the total data and the data excluding the data relating to the crime of financial laundering and usury which were therefore also considered separately. Regarding GDP per-capita, subsequently, we divided it into 4 income bands (0-15000, 15001-28000, 28001-50000, >50000).

- ***17 sustainable development goals***

Last but not least, with reference to the environmental context in which the properties are built, it is necessary to consider all those factors that impact the desirability, buildability of a territory and construction safety. Geology, seismicity, climate, environmental degradation, the risk of landslides and floods, solar radiation are all relevant aspects. Not least in importance are economic and social conditions. To encompass all these aspects, we collected data relating to the indicators proposed by the Eni Enrico Mattei Foundation, and the Italian network of the Sustainable Development Solutions Network (SDSN Italia) to monitor the 17 sustainable development goals in the various Italian cities. These data were available on the site https://sdg-portal.it/it.

We obtained descriptive statistics measures for each indicator, and we provided a graphical representation through a histogram. Each histogram illustrates the distribution of the values assumed by each indicator (Figures 2, 3 and 4).

In the following we reported the goals divided into three environmental, economic, and social pillars. However, the chosen division criterion does not affect the result of our analysis.

1) The environmental pillar:

Goal 2 Zero hunger (*Urban gardens*: how many square meters of decommissioned or abandoned surface area per 100 inhabitants have been converted to the organic cultivation of fruit and vegetables; *Obesity and severe obesity*: the percentage of individuals aged 18+ who are obese or overweight); **Goal 6 Clean water and sanitation** (*Network losses*: the percentage of water losses on the total volume released into the network; *Resident population connected to urban wastewater purification plants* (%); *Resident population served by the urban wastewater sewerage network* (%); **Goal 7 Affordable and clean energy** (*Solar thermal and photovoltaic every 1000 inhabitants*: the amount of solar thermal and photovoltaic power installed on public buildings per 1,000 inhabitants; *Photovoltaic solar panels per km2*: the amount of photovoltaic solar power installed per km2); **Goal 12 Responsible consumption and production** (*Separate waste collection*: the percentage of separate waste out of the total waste produced; *Production of urban waste*: kg of urban waste produced per inhabitant each year; *Residential waste collection*: the percentage of the population served by residential waste collection out of the total population; *Recycling points*: square meters per 100,000 inhabitants designed for ecological islands); **Goal 13 Climate action** (*CO2*: the total CO2 equivalent emitted by each inhabitant; *Population exposed to flood risk*: the percentage of the population exposed to flood risk out of the total population); **Goal 15 Life on land** (*Ecolabel licenses*:

the percentage of companies with the Ecolabel mark out of the total number of companies; *Usable urban greenery*: square meters of urban green space per inhabitant).

The **Goal 14 Life below water** was not considered for comparability reasons.

1) The economic pillar

Goal 1 No poverty (*Economic suffering index*: the percentage of families with an annual income between 0 and €10,000 at the time of the declaration; *Individuals in families with low work intensity*: the percentage of individuals in families with work intensity levels less than 0.20); **Goal 8 Decent work and economic growth** (*The average taxable income per capita*; *NEET between 15 and 29 years*: the percentage of young people aged 15-29 who neither study nor work out of the total number of young people aged 15-29; *Early exit from the education and training system*: the percentage of 18-24 year olds who have completed secondary school and are not currently enrolled in any education or training program); **Goal 9 Industry, Innovation, and Infrastructure** (*Mobility offered by transport public*: number of kilometers per inhabitant covered each year by public transport); **Goal 10 Reduced inequalities** (*GINI index:* the GINI coefficient, which represents the distribution of wealth among residents - most commonly used to measure inequality; *Digital divide*: the percentage of the population excluded from fixed and mobile broadband); **Goal 17 Partnership for the goals** (*Broadband access*: the percentage of households served by a potential internet connection of at least 30 Mbps; *Social cooperatives*: number of social cooperatives per 10,000 inhabitants).

2) The social pillar

Goal 3 Good wealth and well-being (*Life expectancy at birth*: how many years can a newborn baby expect to live assuming that in the future the mortality is constant at each age; *Life expectancy at 65 years*: how many years can a 65-year-old person expect to live assuming that in the future the mortality is constant at each age; *Deaths and injuries in road accidents*: number of deaths or injuries in road accidents per 1,000 inhabitants; *Suicide mortality and intentional self-harm*: the total number of deaths due to suicide or intentional self-harm based on the territory of residence; *Infant mortality ratio*: the total number of children under one year of age who die for every 1,000 live births); **Goal 4 Quality education** (*Index of care of users of childcare services*: the percentage of children under 3 years old who attend childcare services; *Level of alphabetic competence*: the score that students in the second grade of secondary school obtained in the alphabetic proficiency tests; *Level of numerical competence*: the score that students in the second grade of secondary school obtained in numerical competence tests; *Population with a middle school diploma*: the percentage of students who have obtained a secondary school diploma out of the total population; *Population*

0-16 enrolled in the public education cycle: the percentage of individuals aged 0-16 who attend childcare services, first year of primary school or who are enrolled in a regular course of studies out of the total population aged 0-16; *Schools with a ramp*: the percentage of schools equipped with a ramp out of the total number of schools); **Goal 5 Gender equality** (*Absolute difference between male employment rate and female employment rate*: the difference between the percentage of men who work and the percentage of women who work; *Education level of women*: the percentage of women graduates out of the total number of graduates; *Women enrolled in university courses*: the percentage of women enrolled at university out of the total number of students enrolled); **Goal 11 Sustainable cities and communities** (*Cycle paths*: meters of surface used for cycle paths for every 100 inhabitants; *PM2.5 µg/m3*: µg/m3 of PM2.5 recorded on average every year; *Housing quality*: number of individuals living in homes without a toilet per 100,000 inhabitants; *PM10 µg/m3*: amount of µg/m3 of PM10 recorded on average each year; *Noise pollution*: number of noise pollution complaints per 100,000 inhabitants; *Nitrogen dioxide NO2 µg/m3*: amount of µg/m3 of NO2 recorded on average each year; *Pedestrianized road surfaces*: number of square meters of surface per inhabitant used for pedestrian traffic; *Dead, missing and people directly affected by disasters*: number of dead, missing and people directly affected by disasters per 100,000 inhabitants); **Goal 16 Peace, Justice and strong institutions** (*Electoral participation in politics*: the percentage of the population that took part in political elections; *Court efficiency*: average number of days of pending civil proceedings in the year).

In figures 2, 3 and 4 we plot histograms for all the indicators of each pillar. The descriptive statistics show that for the environmental pillar, the indicators characterized by a higher dispersion, computed as the interquartile range are a) Urban gardens (m2); b) Solar thermal and photovoltaic every 1000 inhabitants (kW); c) Population exposed to flood risk (%). For economic pillar the indicators characterized by a higher dispersion is Digital divide (%). Last, but not least, for the social pillar the indicators with a higher dispersion are a) Cycle paths (m); b) Noise pollution (Number); c) Pedestrianized Road surfaces (m2); d) Dead, missing and people directly affected by disasters (Number); e) Population exposed to flood risk (%).

We merged all the previous data to obtain a database of 5500 rows by 61 columns. Then we added two columns to this database relating to the coordinate (longitude and latitude) of the various provinces considered.

Figure 2: Environmental pillar. A histogram for each indicator

Figure 3: Economic pillar. A histogram for each indicator

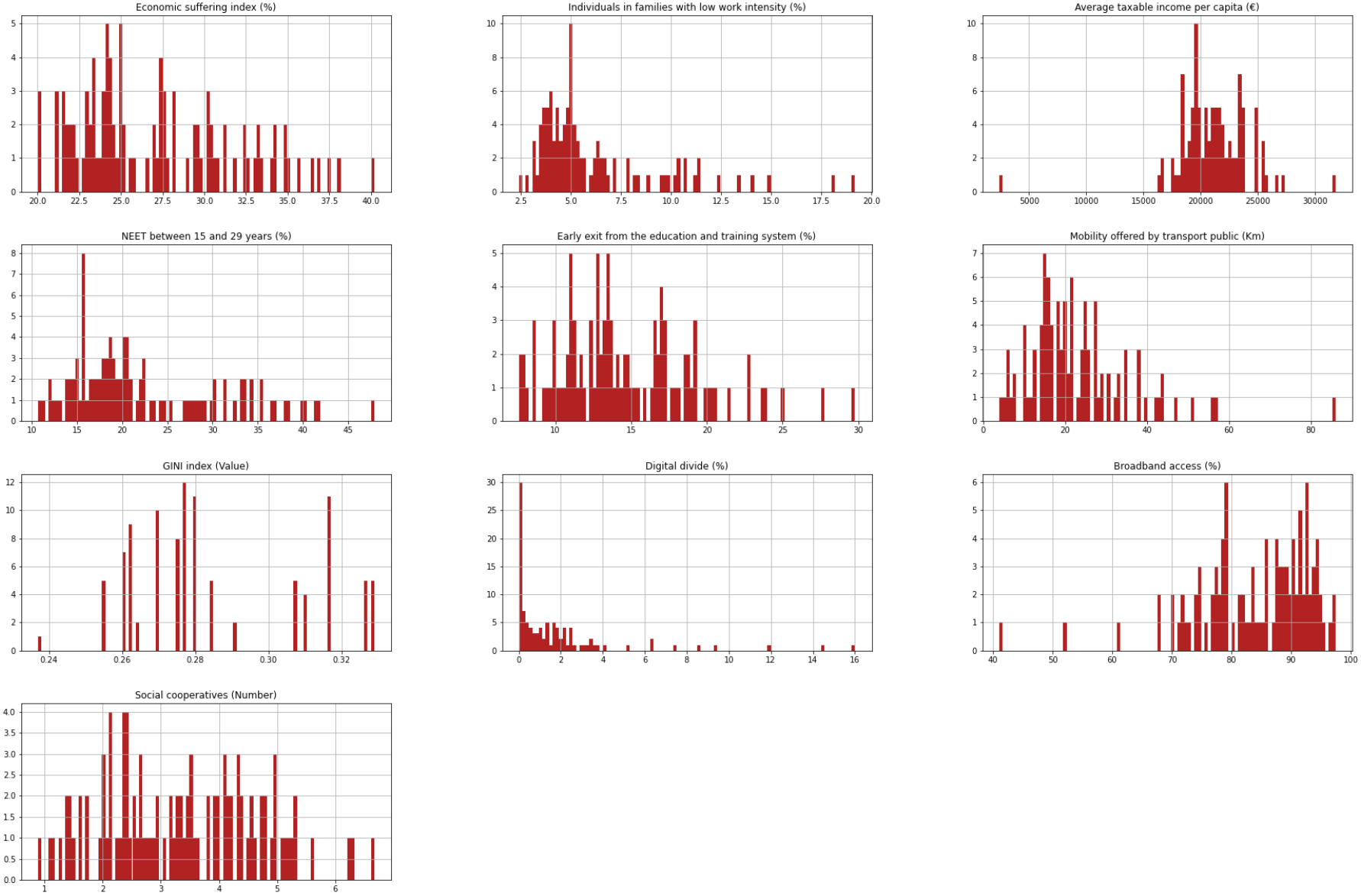
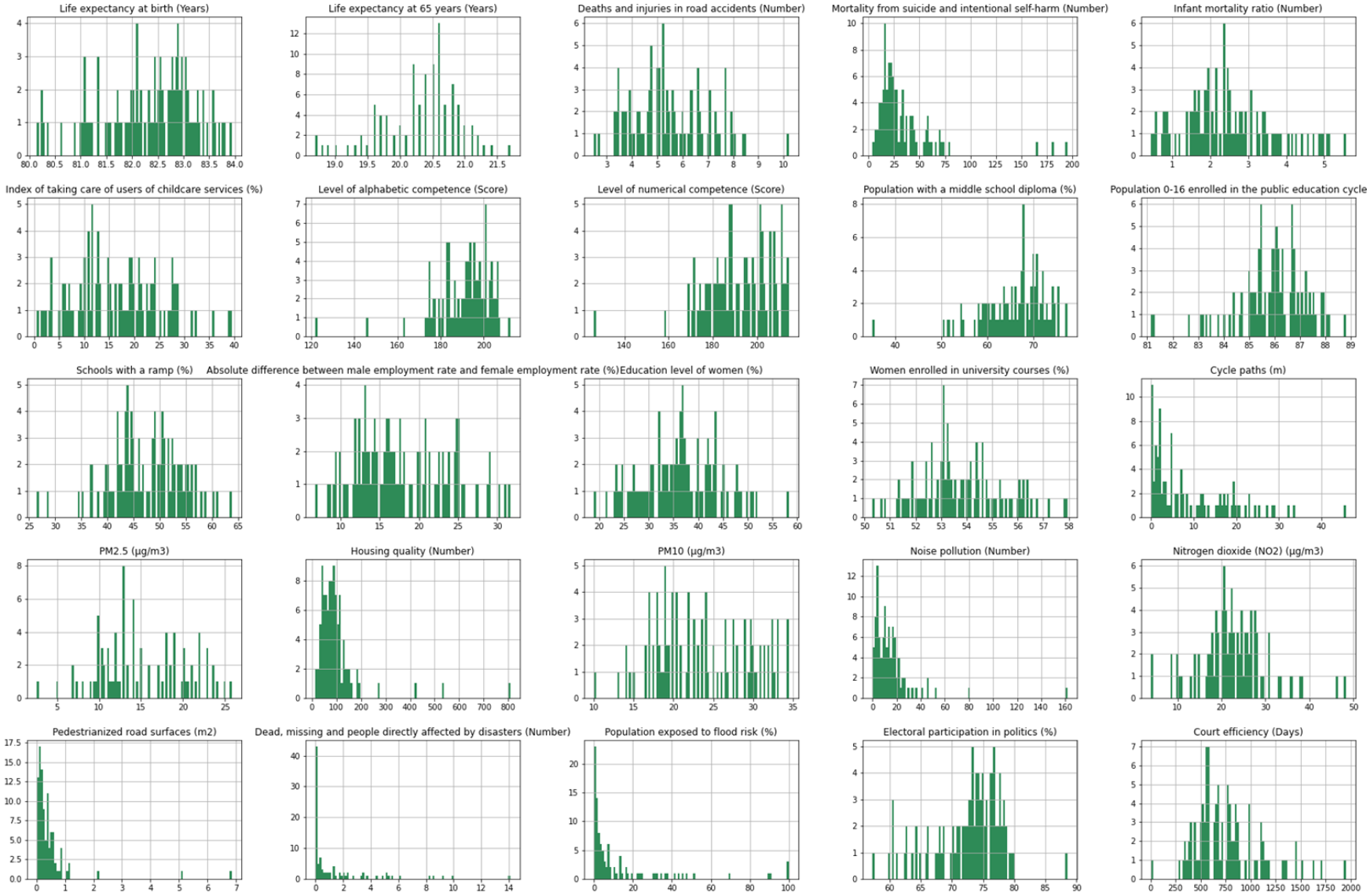


Figure 4: Social pillar. A histogram for each indicator



4. Machine learning model for price prediction and results discussion

Based on the data set obtained in the previous section, we trained a machine learning model for predicting the sales prices of the properties under analysis. This model belongs to the Ensemble Learning algorithms category. Specifically, it is the implementation of Gradient Boosting Trees available within the Scikit-learn Python library. For a theoretical introduction to this category of algorithms, you can see Hastie T, Tibshirani R and Friedman J (2009, 2013).

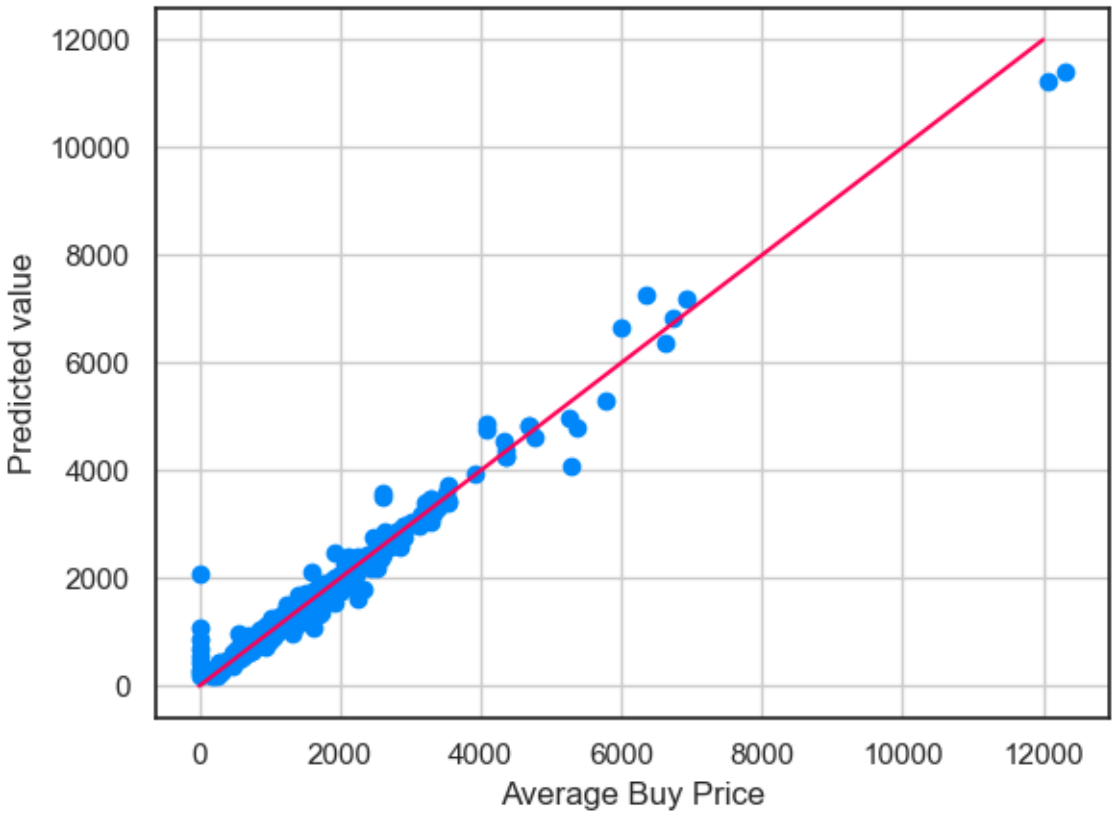
To proceed with the training, we split the data set into two subsets, one for training and one for testing. In the test data set we inserted 15% of the rows of the complete dataset, randomly chosen using a stratification criterion based on the income brackets of GDP per capita. This subdivision was automated via a routine available in the Scikit-learn library. The rest of the rows constituted the training data set.

For the stratification criterion we created a per-capita GDP category attribute with four categories: category one ranges from €0 to €15000, category two ranges from €15000 to €28000, category three ranges from €28000 to €50000, category four includes income higher than €50000.

Since each algorithm has free-to-vary parameters that affect the performance of the model, we used a Scikit learn technique for parameter optimization, to make sure that the model has optimal performance. This technique is called hyper parameter tuning (an explanation is provided here: https://scikit-learn.org/stable/modules/grid_search.html#searching-for-optimal-parameters-with-successive-halving).

Using the parameters identified with this technique, the price values predicted by the model agree with the observed price values (see Figure 5). The goodness of the model measured through the R^2 on the test set is equal to 0.97. The graphical representation of the relationship between these two variables is available in figure 5.

Figure 5: Comparison between predicted and observed prices' value.



4.1. The Shapley

Once we obtained the price prediction model, we asked ourselves which variables, among those present in our data set, had a greater impact on the prediction. To answer this question, we used Python's SHAP library and the test data set as a database.

For each input variable, the explanation algorithm assigns a value based on the contribution that each input variable gives to the prediction of the output variable, i.e. the predicted value of prices (Štrumbelj and Kononenko 2014). The underlying logic, used to assign such values, is cooperative game theory. Specifically, Shapley considers different subsets of the input variables and based on these realigns the explanation model, obtaining the price prediction both in the case in which the input variable whose SHAP value is to be calculated is included, and when this variable is eliminated. Since the effect on the prediction of including and excluding a variable varies as the subset considered varies, for all subsets the difference between the predicted price including the variable and the predicted price excluding the variable is calculated and a weighted average is made (for a theoretical analysis of the SHAP value can be seen Lundberg S.M. and Lee S.I., 2017)

Among the various graphs available for representing the SHAP Value results, we used the bar plot (Figure 6) and the violin plot (Figure 7).

Figure 6 The summary plot to show the feature importance of each feature in the model.

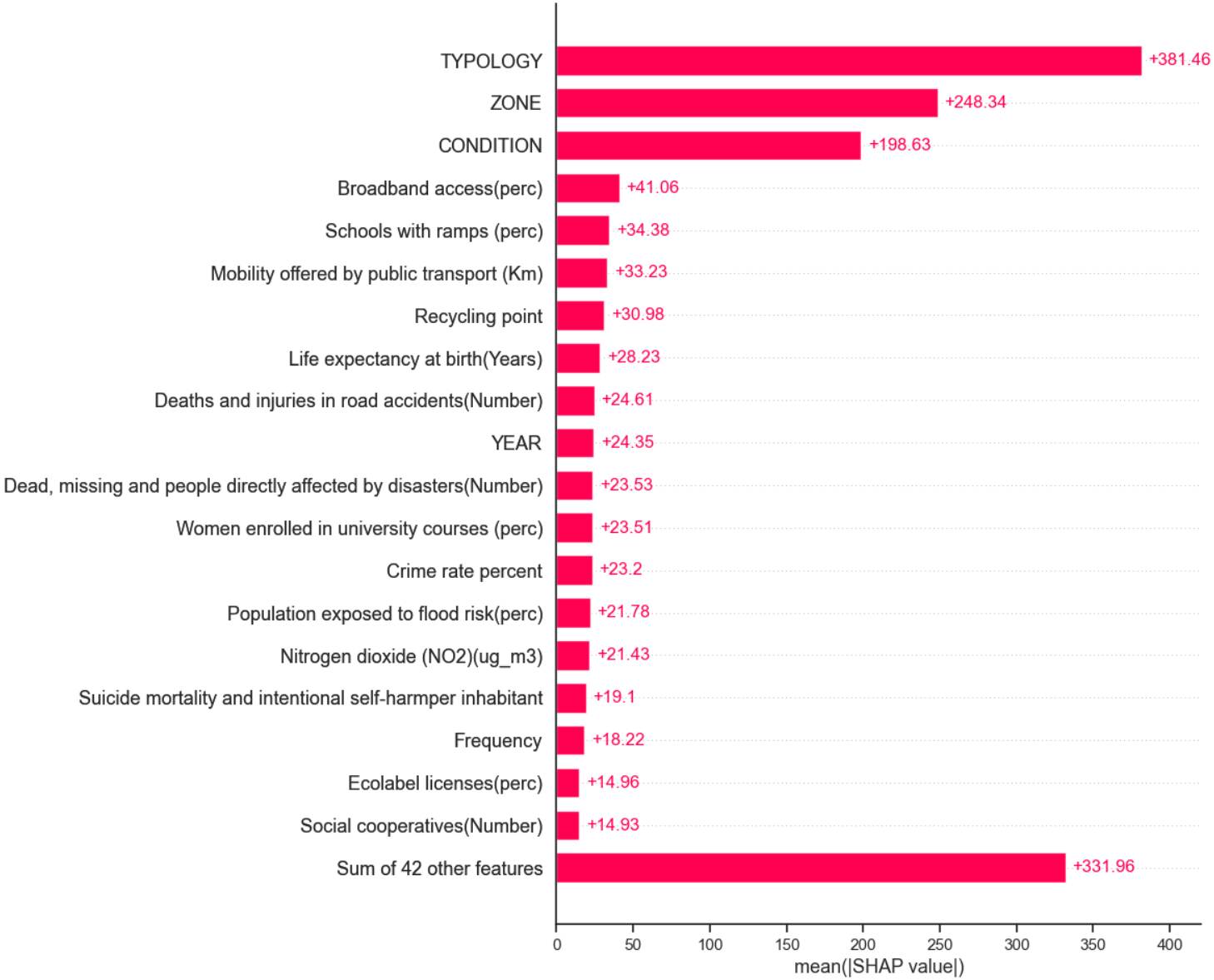
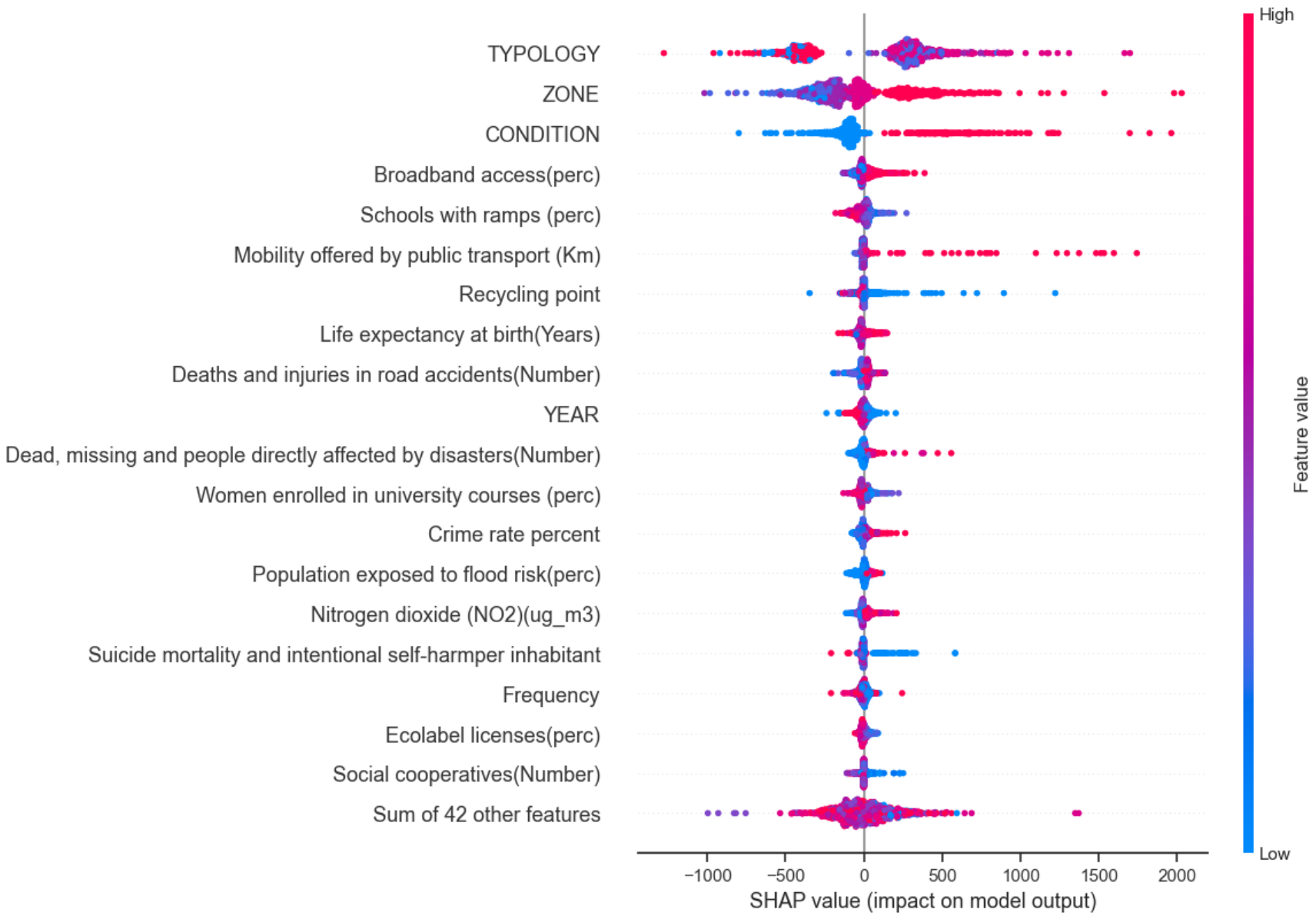


Figure 7: Violin plot to show the most impactful features



In the first graph on the x-axis the range of total variation of the average value of the SHAP is shown for all the input variables; on the ordinate axis we have the input variables sorted in decreasing order by average SHAP value.

In this plot we show that, among the top 20 variables that are important for the prediction of the model, we find the categorical features (zone, typology and condition), 'Broadband access (%) (goal 17)', 'Schools with ramps (%) (goal 4)', 'Mobility offered by public transport (Km) (goal 9)', Recycling point (m2) (goal 12)' , 'Life expectancy at birth (Years) (goal 3)', 'Deaths and injuries in road accidents (Number) (goal3)', 'Women enrolled in university courses (%) (goal 5)', 'Dead, missing and people directly affected by disasters (Number) (goal 11)', 'Crime rate percent', 'Population exposed to flood risk (%) (goal 11)', 'Nitrogen dioxide (NO2) (µg/m3) (goal 11)', 'Suicide mortality and intentional self-harm per inhabitant (goal 3)', 'Frequency', 'Ecolabel licenses (%) (goal 15)', 'Social cooperatives (Number) (goal 17)'. The identified variables belong to all three pillars considered. Therefore, all three aspects affect the value of non-residential properties: environmental, economic, and social.

Turning to figure 7 we observe that the violin plot represents the distribution of the SHAP values assumed by the most relevant input variables. The colour of the various points denotes the value assumed by the input variable. For example, taking a point with a high SHAP value, if the colour of the point is blue it means that the value assumed by the feature (in a specific row of the data set) is low; if it is red, it means that the value assumed by the input variable in the data set is high. The variation range of the SHAP value is therefore placed on the x-axis; the vertical axis shows the input variables in order of importance.

Observing the violin plot we notice that for some variables the colour is not clearly distinguished into red and blue but is distributed in a more heterogeneous way. We went to investigate a possible interaction between that variable and other input variables. In the graphs in Figure 8, 9, 10 and 11, we analyze and comment on some of these interactions.

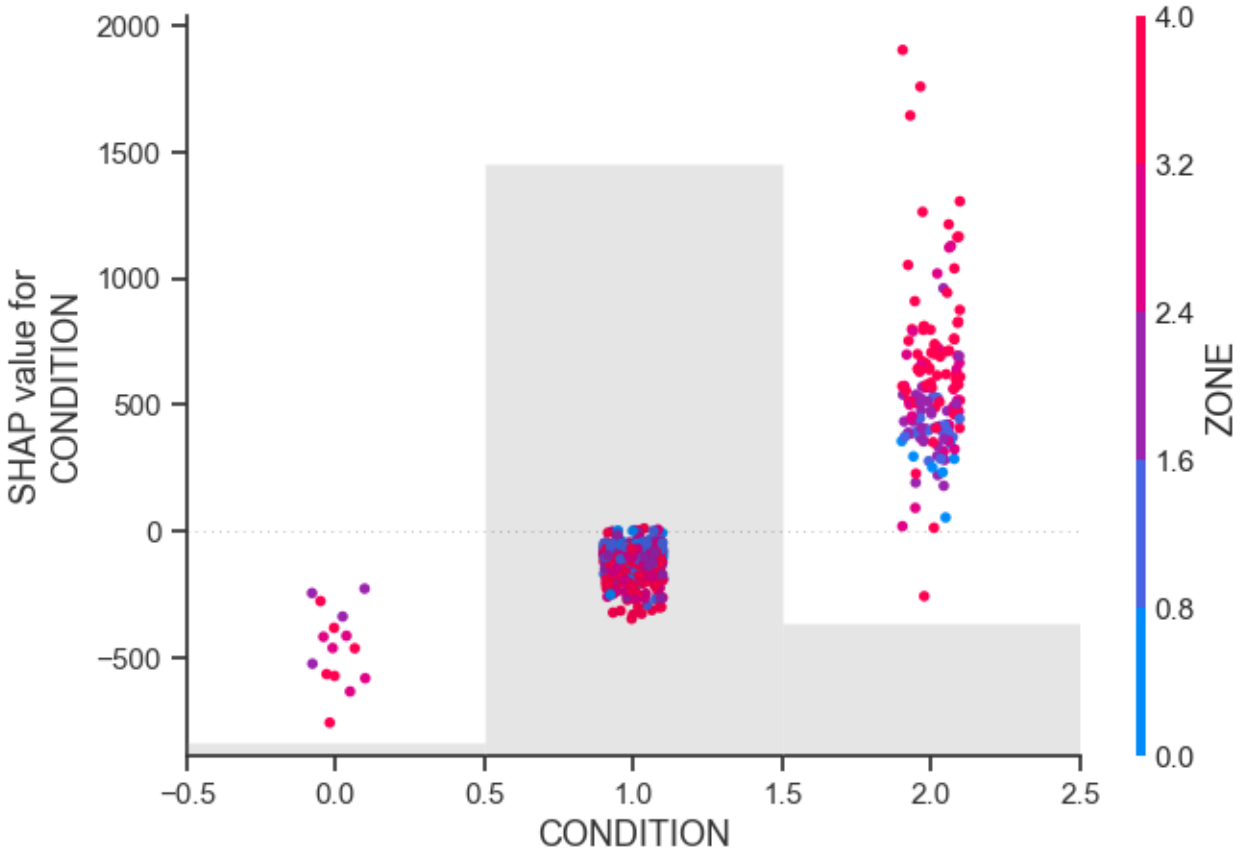
Therefore, going into detail and focusing only on the features that seem most relevant according to the graph in figure 7, we report our results, making a comparison with those obtained from the literature. Some authors studied the relationship between broadband access and house values (Molnar et al., 2019; Wolf and Irwin (2024). Results of such studies show that homes with access to faster broadband connection have larger sale prices. To our best knowledge, there are no studies on the relationship between broadband access and non-residential properties value. With reference to this feature, in Figure 7 we can observe that this variable is mostly high with a positive SHAP value. This means higher broadband access counts tend to positively affect the output.

With reference to recycling points, we observe that a lower value of the feature positively impacts on the model prediction. This result is consistent with those available in the literature. Recent studies suggest that all types of recycle points affect real estate prices, but sites classified as hazardous, especially aquatic hazardous sites, are associated with the lowest real estate prices. Moreover, non-residential properties are more likely to be located near waste sites and thus may reveal greater price impacts (Ihlanfeldt and Taylor 2004; Braden et al., 2011).

Another result that is both evident and predictable is the role played by the mobility offered by public services. We show that a higher value of the ''Mobility offered by public transport' feature is positively linked to our price prediction. Empirical evidence supports this hypothesis both in medium sized city - without mobility problems - and a very large city with congestion problems (Cordera et al., 2019).

Moving on to the categorical variables, we can clearly see the positive relationship with the SHAP value. A better conservation state is associated with higher prices. Looking at the interaction between the feature "condition" and "zone" (see figure 8), we can see that for those properties in excellent condition, localization in central areas produces a positive impact on output. Previous studies have analysed the role of property conditions on home prices (Nanda and Ross, 2012). Miller et al. (2018) find that in favourable market conditions, the price differences determined by the different conservation states of the properties are not particularly relevant. Conversely, if market conditions weaken, the qualitative differences between high-quality properties have a significant impact on prices.

Figure 8: SHAP interaction plot between 'condition' and 'zone'

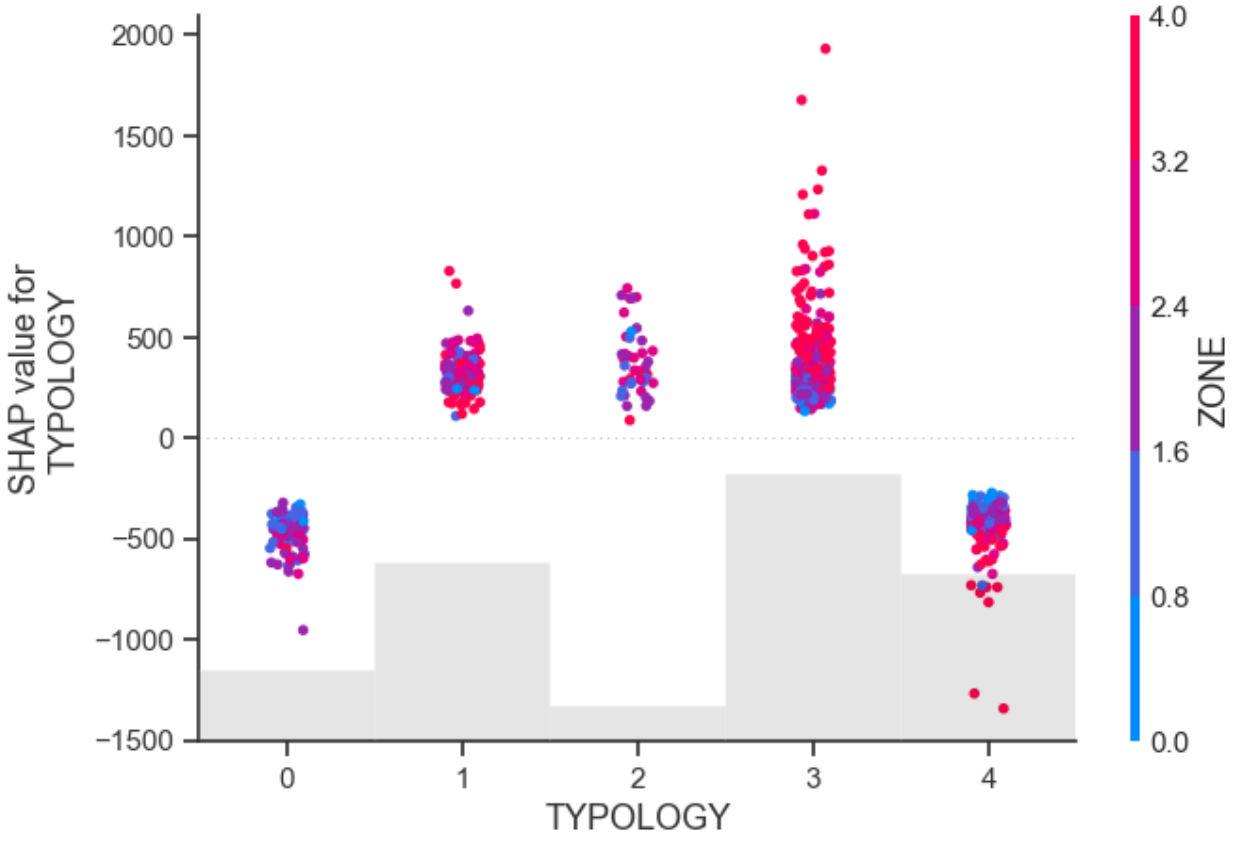


Note.

The zone considered are 0=R; 1=E; 2=D; 3=C; 4=B, where B=centre, C=semi-centre, D=outskirts, E=suburban area, R=extra-urban area/agricultural area. The conditions are 0=POOR; 1=NORMAL; 2=EXCELLENT.

A further result to highlight, with reference to the categorical variables, concerns the typology of properties considered. Among the various analysed non-residential properties, those that seem to influence SHAP values most are offices, shopping centres, shops and laboratories – mainly localized in central area (see Figure 9). The fact that the distributions of SHAP values assumed by the properties belonging to this sector are consistently higher than the SHAP values assumed by warehouse and industrial sheds, indicates a direct association of these features with prices.

Figure 9: SHAP interaction plot between 'typology' and 'zone'



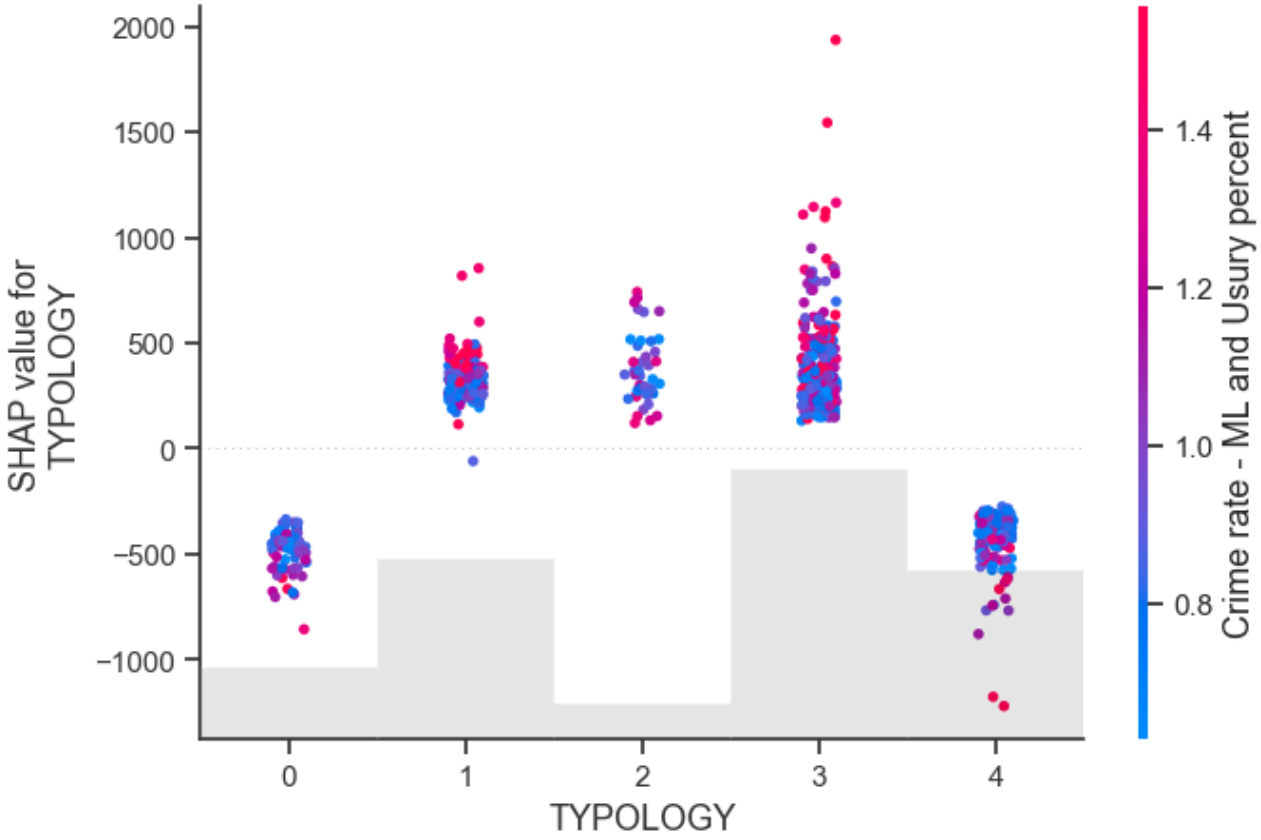
Note.

The typologies considered are 0=INDUSTRIAL SHEDS; 1=OFFICES; 2=SHOP

PING CENTERS; 3= SHOPS AND LABORATORIES; 4= Warehouses

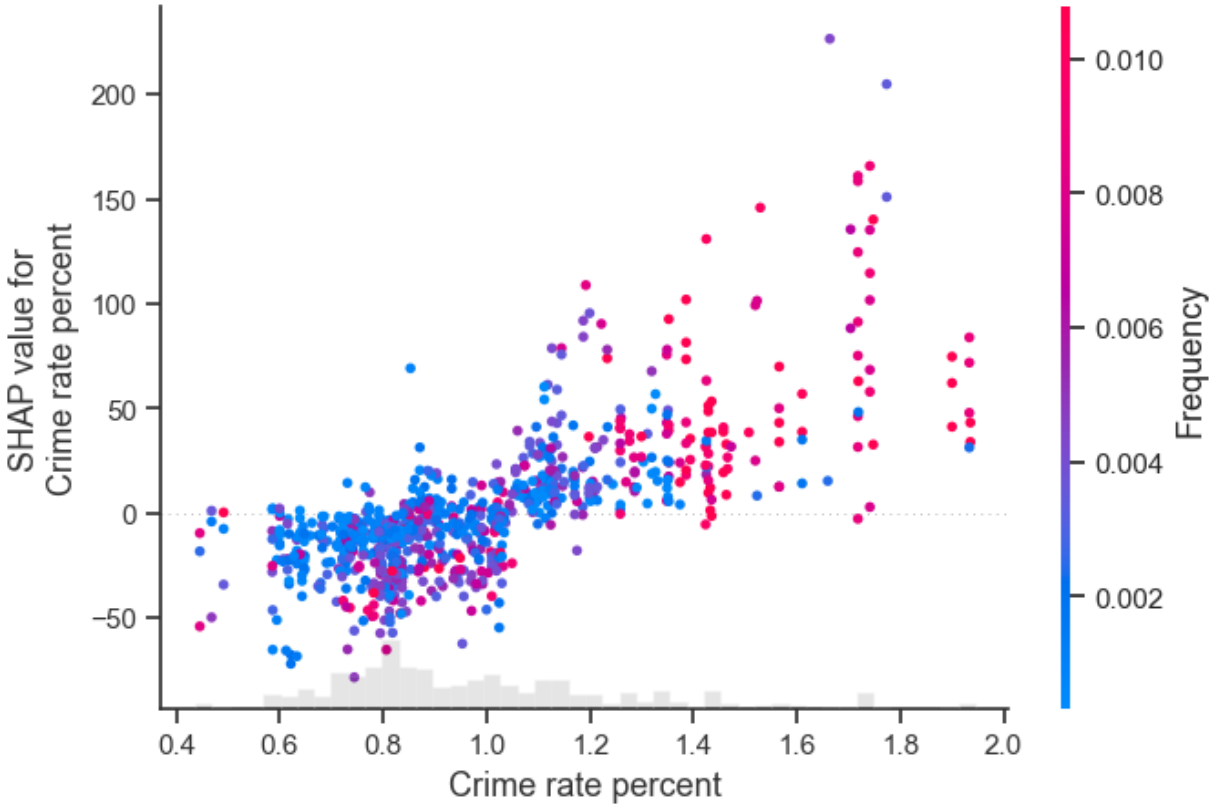
The effect of the interaction with crime rate can be appreciated in Figure 10 -- especially for offices, shops and laboratories -- by the increasing vertical distinction in blue (low crime rate)/red (high crime rate) colours. We can observe that the SHAP value of these properties is higher in coincidence with high crime rates.

Figure 10: SHAP interaction plot between 'typology' and 'crime rate minus ML and usury'



Moreover, in figure 11 we see an increasing relationship between crime rate and transaction frequency. Higher crime rates correspond to higher SHAP values. On the contrary, lower crime rates correspond to low SHAP values. At high crime values there is a greater frequency of real estate transactions. The relationship between crime and property value has been analysed in the literature, with reference to the residential market. The main results found a positive or negative relationship depending on the type of crime considered. On one side, acts of vandalism and other forms of violence committed by criminals have the effect of reducing housing prices, spreading a climate of fear and insecurity (Dugato et al., 2015; Kruisbergen et al., 2015). On the other side, the impact of money laundering on the real estate sector goes in the opposite direction (Barone, 2023). The most common forms in which the proceeds of crime entered the real estate market involved mortgages and cash or private digital currencies as close substitutes (Schneider, 2004; Borgonovo et al., 2021).

Figure 11: SHAP interaction plot between ‘crime rate and ‘frequency’

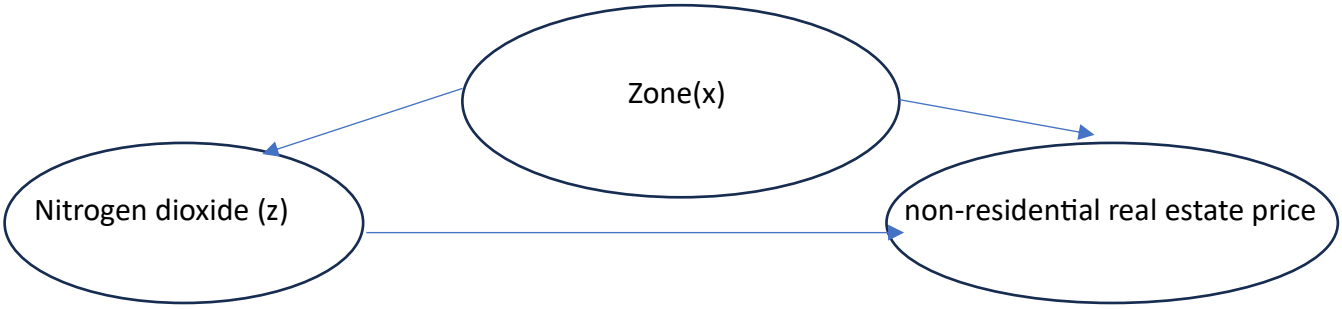


4.2. The causal inference

Having determined those which, according to the SHAP criterion, are the variables that have the greatest impact on the results, i.e. on the determination of the output variable, we wonder whether some of these variables may in fact not have a direct effect on the output variable, but rather it happens that these impact on other input variables which are instead directly linked to the output. To answer this question, we need to proceed with causal inference.

To perform this analysis, we used the DoWhy and EconML libraries from the PyWhy project (Sharma, 2020). To test our hypothesis, we adopted the back-door criterion implemented by the double machine learning (DML) method from EconML. The aim of DML is to estimate (heterogeneous) treatment effects in case of a lot of observed confounders (Chernozhukov2016).

Focusing on the variables that are relevant according to the SHAP values, we hypothesize that those that effectively impact the output (i.e. the price of properties) are the crime rate (%), the broadband access (%), Nitrogen dioxide concentration (NO2) (µg_m3), recycling points (m2), ecolabel licenses (%) and mobility offered by public transport (Km). The idea is that these variables, more than others, may be controlled by the government so they can directly impact on the output. As confounders we considered 'zone', typology', 'condition', 'social cooperatives(number)', 'life expectancy at birth(years)', 'women enrolled in university courses (perc)', 'dead, missing and people directly affected by disasters(number)', 'suicide mortality and intentional self-harmer inhabitant', 'deaths and injuries in road accidents(number)', 'schools with ramps (perc)', 'frequency', 'Population exposed to flood risk(perc)'. In figure 12 we provide the cause-and-effect graph. However, given the numerous variables, the graph is not particularly explanatory. Nevertheless, in a nutshell, the graph is composed of nodes and arrows. The nodes are the variables considered and the arrows allow the relationships between the variables to be established. Below I illustrate an example of reduced representation, where a confounders variable (which we define as x) impacts a treatment variable (z) which instead has a direct effect on the output (y). The effect of the confounders on the output variable is therefore only an indirect effect.



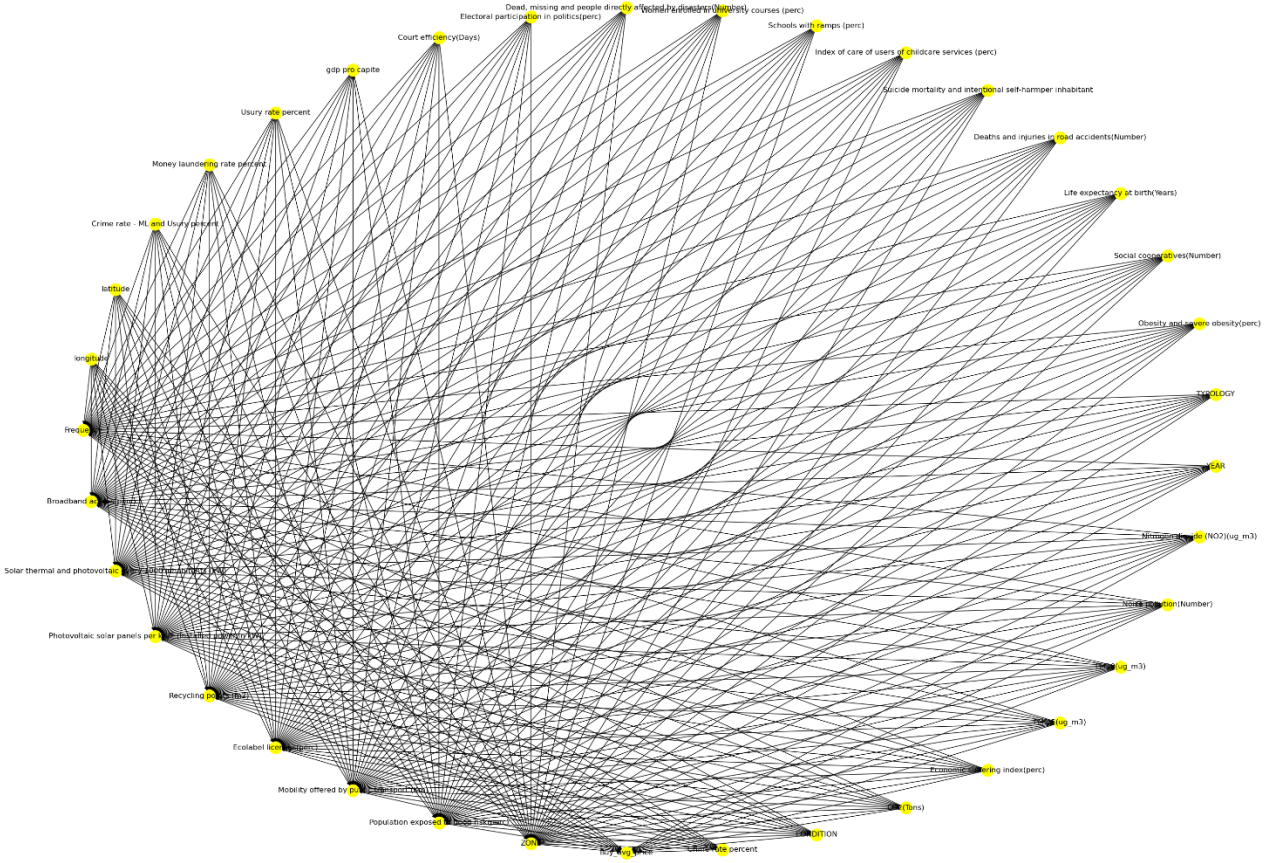
Our estimate finds a mean treatment effect equal to 449.86 €/sqm with a 95% confidence interval between 320.95 and 533.36 €/sqm. To assess the robustness of the estimate, we used the “add a random common cause” as a refute test. With this test we suppose the existence of an independent random variable as confounder that is correlated with the treatment and outcome, and we add it to the model. If the model is well designed to control for confounders, the causal estimates should not change so much. While the estimated effect was 449.86, the new effect with the refutation test is 453.06, with a p-value of 0.16. Therefore, since the p-value is greater than the selected threshold of 0.05, we cannot reject the null hypothesis.

5. Conclusions

In this paper we propose a machine learning approach to analyse the determinants of non-residential real estate market prices. To predict the prices, we used a gradient boosting regressor trained on a custom dataset composed of both data coming from the Real Estate Market Observatory (OMI) of the Italian Revenue Agency and data encompassing, for each given Italian province, the 3 main pillar of sustainability: environmental, social, and economic. The best predictive model shows a $R^2 = 0.972$. Then we evaluated the feature importance and to explain the result of the model we used the Shapley Additive exPlanations (i.e. SHAP) approach and causal inference methods.

It results that all three pillars (social, environmental, and economic) impact on non-residential properties price. The most important goals are Goal 3 (Good wealth and well-being); Goal 11 (Sustainable cities and communities); Goal 12 (Responsible consumption and production); Goal 17 (Partnership for the goals). Among the most important variables, out of the pillars, we find the categorical variables, i.e. location of the properties, typology, and conservation state. For some typology of properties (offices, shops, and laboratories) that have the greatest positive impact on prices, the crime rate would seem to play a key role. Precisely, we can observe that the price of these properties is higher in correspondence of higher crime rates. However, the SHAP values allowed us to assess the correlation between variables. Correlation and causation are two different issues. To find out which of the variables have an actual causal relationship with the price of properties, we performed a causal inference analysis. In this analysis we hypothesized that the variables that have a causal relationship with prices are those that can be controlled by policy makers more than others. Our hypothesis was confirmed by the data.

Figure 12:Causal graph that show the causal relatiosìnship between features and outcome



References

- Agarwal S, Ambrose BW, Lopez LA, Xiao X (2020) Did the paycheck protection program help small businesses? Evidence from commercial mortgage-backed securities. Health & the Economy eJournal. doi:10.2139/ssrn.3674960 Corpus ID: 234650284
- Barone R (2023) Home sweet home, how money laundering pollutes the real estate market: an agent based model. Journal of Economic Interaction and Coordination 18, 779–806. https://doi.org/10.1007/s11403-023-00391-y
- Batalha M, Goncalves D, Peralta S, dos Santos JP (2022) The virus that devastated tourism: The impact of covid-19 on the housing market. Regional Science and Urban Economics 103774. ISSN 0166-0462, in press, available online. https://doi.org/10.1016/j.regsciurbeco.2022.103774
- Baptista, R., Farmer, J. D., Hinterschweiger, M., Low, K., Tang, D., & Uluc, A. (2016). Macroprudential policy in an agent-based model of the UK housing market. Staff Working Paper 619, Bank of England.
- Beland LP, Brodeur A, Wright T (2020) COVID-19, stay-at-home orders and employment: Evidence fromCPS data. IZA DP No. 13282.
- Borgonovo E, Caselli S, Cillo A, Masciandaro D, Rabitti G (2021) Money, privacy, anonymity: What do experiments tell us?. Journal of financial stability, 56. https://doi.org/10.1016/j.jfs.2021.100934
- Bowman R and Wills J (2008) Valuing Green: How Green Buildings Affect Property Values and Getting the Valuation Method Right, Australian Green Building Council, Melbourne.
- Braden J B, Feng X and Won D (2011) Waste Sites and Property Values: A Meta-Analysis. Environ Resource Econ 50, 175–201. https://doi.org/10.1007/s10640-011-9467-9
- Breiman L (2001) Random Forests, Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
- Capponi A, Rios DA (2020) COVID-19 mortgage forbearance: Implications on the housing market. SSRN Electronic Journal 3618776.
- Carro A, Hinterschweiger M, Uluc A, Farmer J D (2022) Heterogeneous effects and spillovers of macroprudential policy in an agent-based model of the UK housing market. Industrial and Corporate Change. dtac030, https://doi.org/10.1093/icc/dtac030
- Chen L, Yao X, Liu Y, Zhu Y, Chen W, Zhao X, and Chi T (2020) Measuring impacts of urban environmental elements on housing prices based on multisource data—A case study of

Shanghai, China. International Journal of Geo-Information, 9(2), p. 106. https://doi.org/10.3390/ijgi9020106

- Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, and. Newey W (2016) Double Machine Learning for Treatment and Causal Parameters. ArXiv e-prints, July 2016.
- Cincotti S, Raberto M and Teglio A (2022) Why do we need agent-based macroeconomics?. Review of Evolutionary Political Economy 3, 5-29. https://doi.org/10.1007/s43253-022-00071-w
- Cordera R, Coppola P, Dell' Olio L, Ibeas A (2019) The impact of accessibility by public transport on real estate values: A comparison between the cities of Rome and Santander, Transportation Research Part A: Policy and Practice, Volume 125: 308-319. ISSN 0965-8564, https://doi.org/10.1016/j.tra.2018.07.015.
- De Villa A R (2023) Causal Inference for Data Science. MEAP. ISBN 9781633439658
- D'Lima W, Lopez LA, Pradhan A (2021) COVID-19 and housing market effects: Evidence from U.S. shutdown orders. Real Estate Economics 50:303–339. https://doi.org/10.1111/1540-6229.12368
- Dave D, Friedson A, Matsuzawa K, Sabia J (2020) When do shelter-in-place orders fight COVID-19 best? Policy heterogeneity across states and adoption time (Technical Report w27091). National Bureau of Economic Research. Cambridge, MA. Retrieved fromhttp://www.nber.org/papers/w27091.pdf
- De Liso N (2023) Artificial Intelligence and Inequality. In Global Handbook of Inequality Jodhka, AG, Rehbein (eds.), Springer Nature Switzerland. 2023 S. S. B, https://.org/10.1007/978-3-030-97417-6_49-1.
- Duca JV, Hoesli M, Montezuma J (2021) The resilience and realignment of house prices in the era of COVID-19. Journal of European Real Estate Research 14(3):421–431. https://doi.org/10.1108/JERER-11-2020-0055
- Dugato M, Savarin S, Giommoni L (2015) The Risks and Rewards of Organized Crime Investments in Real Estate. The British Journal of Criminology 55(5):944–965. https://doi.org/10.1093/bjc/azv002
- Franco and Cutter (2022) The determinants of non-residential real estate values with special reference to environmental local amenities. Ecological Economics, 201. https://doi.org/10.1016/j.ecolecon.2022.107485

- Fuerst F and Mcallister P (2008a) “Green noise or green value? Measuring the price effects of environmental certification in commercial buildings”, School of Real Estate and Planning, Henley Business School, University of Reading, Reading, MA.
- Fuerst F and Mcallister P (2008b) “Pricing sustainability: an empirical investigation of the value impacts of green building certification”, working paper from the Proceedings of the American Real Estate Society Conference, April, Florida, ARES.
- Geanakoplos J, Axtell R, Farmer JD, Howitt P, Conlee B, Goldstein J, Hendrey M, Palmer NM, and Yang CY (2012). Getting at Systemic Risk via an Agent-Based Model of the Housing Market. American Economic Review, 102 (3): 53-58. https://doi.org/10.1257/aer.102.3.53
- Granja J, Makridis C, Yannelis C, Zwick E (2020) Did the pay check protection program hit the target?. National Bureau of Economic Research. Series: Working Paper Series n. 27095. Retrieved from http://www.nber.org/papers/w27095
- Griliches Z (1961) Hedonic price indexes for automobiles: An econometric of quality change, Price Statistics of the Federal Government, 173-196, NBER.
- Guenster and Koegst (2018) Environmental responsibility and firm value, in Toine Spapens, Rob White, Daan van Uhm and Wim Huisman (Eds), Green Crimes and Dirty Money, Routledge, London and New York, ISBN: 978-0-8153-7221-9 (hbk); ISBN: 978-1-351-24574-6 (ebk)
- Gupta A, Mittal V, Peeters J, Van Nieuwerburgh S (2022) Flattening the curve: Pandemic Induced revaluation of urban real estate. Journal of Financial Economics, 146(2), 594-636, ISSN 0304-405X, https://doi.org/10.1016/j.jfineco.2021.10.008.
- Hastie T, Tibshirani R and Friedman J (2013) The elements of statistical learning, Springer, Stanford, California (chapter 16) https://hastie.su.domains/Papers/ESLII.pdf
- Hastie T, Tibshirani R, Friefidman J (2009). The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer Series in Statistics, Second Edition, New York, USA. ISBN: 978-0-387-84857-0.
- Hill R J (2013) Hedonic Price Indexes For Residential Housing: A Survey, Evaluation And Taxonomy, Journal of Economic Surveys, 27(5), pp. 879-914. doi: 10.1111/j.1467-6419.2012.00731.x
- Klassen R and McLaughlin C (1996) The Impact of Environmental Management on Firm Performance. Management Science, 42, pp. 1199-1214. https://doi.org/10.1287/mnsc.42.8.1199

- Klügl F, Bazzan ALC (2012) Agent-Based Modeling and Simulation. Association for the Advancement of Artificial Intelligence, AI Magazine: 29-40. ISSN: 0738-4602.
- Kruisbergen EW, Kleemans ER, Kouwenberg RF (2015) Profitability, power, or proximity? Organized crime offenders investing their money in legal economy. *European Journal on Criminal Policy and Research* 21(2): 237–256.
- Ihlanfeldt KR, Taylor LO (2004) Externality effects of small-scale hazardous waste sites: evidence from urban commercial property markets. J Environ Econ Manage 47:117–139 https://doi.org/10.1016/S0095-0696(03)00070-6
- Lancaster K. (1966) A new approach to consumer theory, Journal of Political Economy, 74 (2), pp. 132-157. https://doi.org/10.1086/259131
- Liu S, Su Y (2021) The impact of the COVID-19 pandemic on the demand for density: Evidence from the US housing market. Economics Letters, 207, 110010.
- Liu Y, Tang Y (2021) Epidemic shocks and housing price responses: Evidence from China's urban residential communities. Regional Science and Urban Economics 89, 103695. ISSN 0166-0462. https://doi.org/10.1016/j.regsciurbeco.2021.103695.
- Lucas R E B (1975) Hedonic Price Functions, Economic Inquiry, 13(2), pp.157-178. https://doi.org/10.1111/j.1465-7295.1975.tb00985.x
- Lundberg S.M. and Lee S.I., 2017, A Unified Approach to Interpreting Model Predictions, 31st Conference on Neural Information Processing Systems, available online at https://arxiv.org/pdf/1705.07874
- Mangialardo A, Micelli E, Saccani F (2019). Does Sustainability Affect Real Estate Market Values? Empirical Evidence from the Office Buildings Market in Milan (Italy). Sustainability, 11(12); doi:10.3390/su11010012
- Miller N, Vivek S and Sklarz M (2018) Estimating Property Condition Effect on Residential Property Value: Evidence from U.S. Home Sales Data, Journal of Real Estate Research, 40:2, 179-198. DOI: 10.1080/10835547.2018.12091497
- Molnar G, Savage S J and Sicker D C (2019) High-speed Internet access and housing values, Applied Economics, 51:55, 5923-5936, DOI: 10.1080/00036846.2019.1631443
- Mooya M M (2016) Real Estate Valuation Theory: A Critical Appraisal, Springer-Verlag, Berlin Heidelberg 2016, ISBN 978-3-662-49163-8 ISBN 978-3-662-49164-5 (eBook). DOI 10.1007/978-3-662-49164-5

- Mooya, M M (2022) Property Valuation, Transaction Prices and Market Activity. In: D'Amato, M., Coskun, Y. (eds) Property Valuation and Market Cycle. Springer, Cham. https://doi.org/10.1007/978-3-031-09450-7_17
- Mullainathan S and Spiess J (2017) Machine Learning: An Applied Econometric Approach. Journal of Economic Perspectives, 31 (2): 87-106. DOI: 10.1257/jep.31.2.87.
- Nanda A and Ross S L (2012) The Impact of Property Condition Disclosure Laws on Housing Prices: Evidence from an Event Study Using Propensity Scores. J Real Estate Finan Econ 45, 88–109. https://doi.org/10.1007/s11146-009-9206-y
- Ozel B, Nathanael RC, Raberto M, Teglio A, Cincotti S (2019) Macroeconomic implications of mortgage loan requirements: an agent-based approach. Journal of Economic Interaction and Coordination 14: 7-46. https://doi.org/10.1007/s11403-019-00238-5
- Pearl J (1995) Causal diagrams for empirical research. Biometrika, 82(4), 669–688. https://doi.org/10.2307/2337329
- Pearl J, Glymour M and Jewell N P (2016) Causal inference in statistics: A primer. Chichester, England: John Wiley & Sons.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. (2011) Scikit-learn: Machine learning in Python. Journal of machine learning research:12(Oct):2825–30.
- Porter M E and van der Linde C (1995) Toward a New Conception of the Environment-Competitiveness Relationship. Journal of Economic Perspectives, 9 (4), pp. 97–118. DOI: 10.1257/jep.9.4.97
- Potrawa T and Tetereva A (2022) How much is the view from the window worth? Machine learning-driven hedonic pricing model of the real estate market, Journal of Business Research, 144, pp.50-65 https://doi.org/10.1016/j.jbusres.2022.01.027
- Rosen S (1974) Hedonic prices and implicit markets: Product differentiation in pure competition, Journal of Political Economy, 82(1), pp. 35–55. https://doi.org/10.1086/260169
- Rubin D B (2005) “Causal Inference Using Potential Outcomes: Design, Modeling, Decisions”. In: Journal of the American Statistical Association 100(469), pp. 322–331. https://www.jstor.org/stable/27590541
- Schneider S (2004) Organized crime, money laundering, and the real estate market in Canada. Journal of Property Research 21(2): 99-118. DOI: 10.1080/0959991042000328801
- Sharma A and Kiciman E (2020) DoWhy: An End-to-End Library for Causal Inference. 2020. https://arxiv.org/abs/2011.04216

- Spapens T, White R, van Uhm D and Huisman W (Eds), Green Crimes and Dirty Money, Routledge, London and New York, ISBN: 978-0-8153-7221-9 (hbk); ISBN: 978-1-351-24574-6 (ebk)
- Splawa-Neyman J, Dabrowska D M, and Speed T P (1990) “On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.” In: Statistical Science 5(4), pp. 465–472. https://www.jstor.org/stable/2245382
- Štrumbelj E, Kononenko I (2014) Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst 41, 647–665 https://doi.org/10.1007/s10115-013-0679-x.
- Wolf D and Irwin N (2024) Is it really bridging the gap? Fiber internet's impact on housing values and homebuyer demographics. Journal of Regional Science, 64, 238–271. https://doi.org/10.1111/jors.12670
- Zhao Y (2020) US housing market during COVID-19: Aggregate and distributional evidence. IMF Working Paper No: WPIEA2020212. ISBN/ISSN: 9781513557816/1018-5941. https://doi.org/10.1016/j.econlet.2021.110010