Lion's Den

ING Risk Modelling Challenge

Lion’s Den

Powered by

Residential Real Estate Pricing Model Design

Pre-selection assignment

Flying Vietnamese:

Dat Thai

Giang Nguyen

Ky Anh Le

Trung Nguyen

TABLE OF CONTENT

A. Project overview

Predicting house prices is crucial for banks like ING and the credit industry as a whole because it helps assess mortgage risks, determine loan-to-value ratios, and optimize lending strategies. Accurate price predictions enable banks to make informed lending decisions, minimize default risks, and maintain financial stability. Moreover, in the credit industry, property valuation plays a key role in securitization and investment decisions.

We chose to focus on predicting house prices in Poland, specifically in Warsaw, because it is the country's economic hub with a dynamic real estate market. Warsaw's housing prices are influenced by various factors such as urban development, infrastructure, and economic growth, etc. making it an ideal case study. Additionally, limiting the prediction scope to one city ensures higher model accuracy and relevance by reducing variability across different regions.

The project is implemented using Python language with popular libraries such as pandas, numpy used to manipulate data and scikit-learn to build machine learning models, libraries such as matplotlib and seaborn are also used to visualize data.

During this project, we successfully developed a predictive model for housing prices in Warsaw, incorporating various factors such as location, amenities, and economic indicators. We experimented with multiple machine learning models, including Linear Regression, Random Forest, and XGBoost, etc. achieving the best results with Random Forest, which provided a balance between accuracy and generalizability.

We have faced numerous challenges during the process of generating a suitable dataset and the model inputs. The research around the real estate market has shown us the lack of access to the wider availability of data. Unfortunately, the dataset has been limited to the year 2021 that may lead to outdated patterns and restriction of the research. We have also encountered difficulties with finding other relevant variables including the market trends and ESG factors that had to be carefully selected and manipulated to ensure the fitness of the data since we have narrowed down the regional-level Warsaw. The factors have been found from reliable sources including the government statistics and reports. The synthetic data has been used extensively to cover a larger scope of the factors that can potentially drive the model for predicting the house prices. GenAI is in use to facilitate the process of obtaining the data.

B. Research and Model Design

1. Topic Research

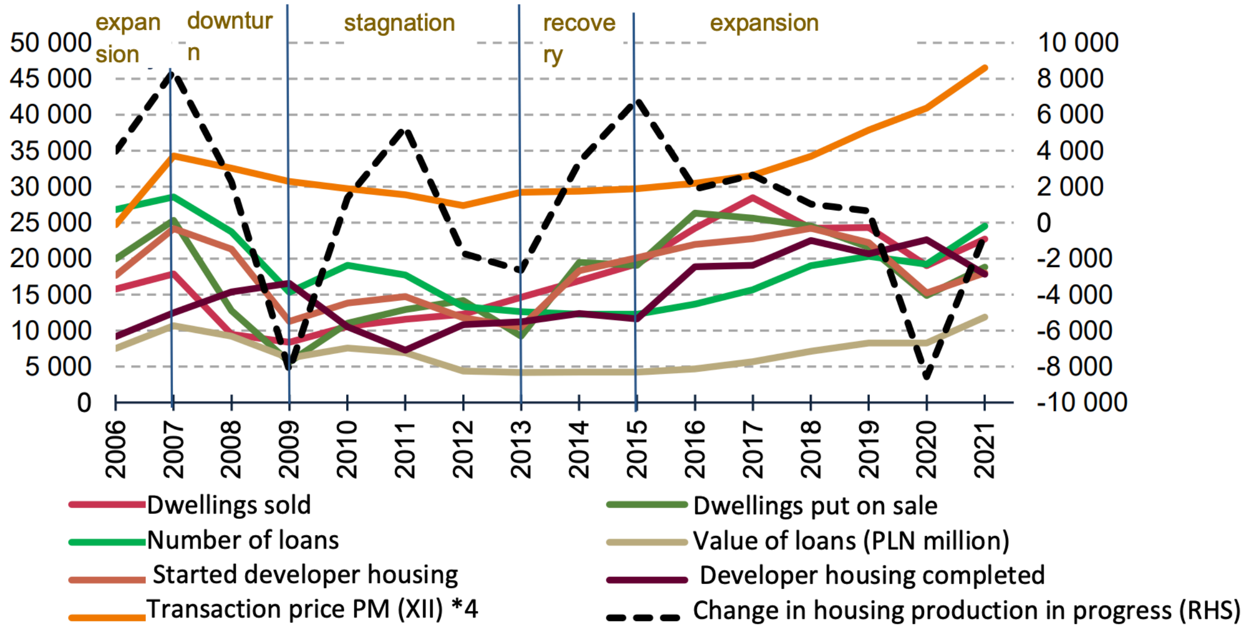
We have approached the topic with our prior understanding of the real estate market combining with the secondary research. The research has been conducted initially for various geographical levels that ranges from continental level such as Asia and Europe to a more national and regional scale within Poland considering the suitable features. Coming across Warsaw with comparison to other cities, we have observed the potential that can be assessed further for the market trends and decided on deep diving into the research of the factors around the capital city. We have analysed the real estate market within Poland and Warsaw since the markets vary significantly between countries due to the differences in e.g. government regulations, interest rates and tax policies. At the research stage, we have prioritised the understanding of the market trends and real estate factors before implementing it into the model inputs and dataset. By having the high-level knowledge of the market, we will be able to assess the viability of the model and highlight key importance of the outputs.

Market cycles shown below in the figure are the structural features of the real estate market. In the residential real estate market, the tension of the market is measured by the common indicator addressing the relationship between the prices and household income. According to the survey from NBP, the residential real estate market remained in the expansion phase, driven by both investment and consumer demand, despite emerging supply-side constraints. While rising home prices and financing costs began to slow market activity in the latter half of the year, demand remained strong, particularly in major cities.

On the supply side, rising construction costs, labor shortages, and limited land availability placed upward pressure on home prices. However, despite these challenges, developers continued to launch new projects, encouraged by high estimated profitability. The financial stability of construction companies and developers remained strong compared to other industries, ensuring continued market activity.

Although demand for housing persisted, affordability concerns emerged, particularly in Poland's largest cities, where rising home prices and increasing interest rates began to limit access to mortgage financing. The probability of a shift in the housing cycle increased, particularly as fiscal and monetary interventions that had extended the expansion phase began to taper off.

Overall, the residential sector in 2021 demonstrated resilience, with high transaction volumes and strong investment interest. However, growing financing costs, supply-side constraints, and economic uncertainty signaled potential shifts in the market cycle in the coming years.



Source: NBP, Statistics Poland, JLL (formerly: REAS), BIK

Figure 1: Housing market cycles in Warsaw

2. Model Input Research

Based on prior research articles, academic papers, especially according to the Hedonic Pricing Models, we can see that both property specific features and external factors play key roles in real estate pricing.

We also perform the feature importance extract method on some real estate pricing datasets, using different models to investigate what type of feature is a must to a house price prediction dataset.

Dataset 1: Real Estate Listings in Portugal

- Feature Importance results (Random Forest Regressor model):

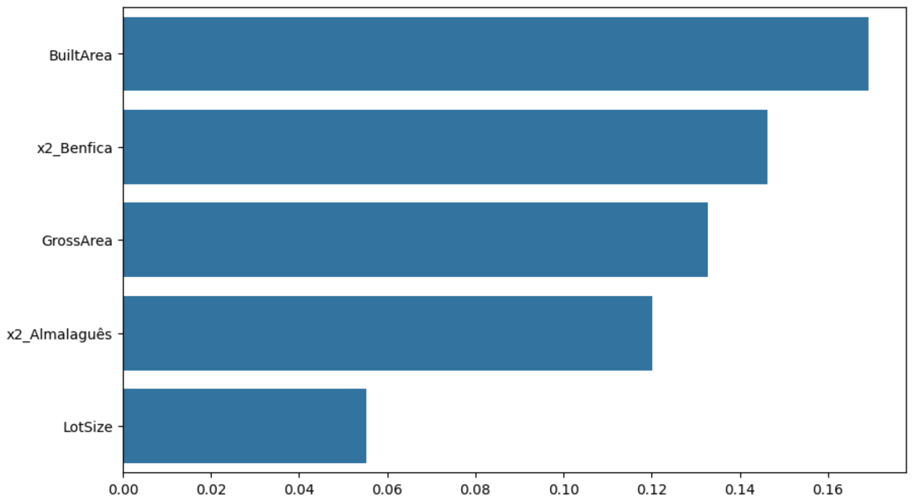


Figure 2: Feature Importance barplot of Real Estate in Portugal using Random Forest Regressor model

Conclusion: The overall pattern of house prices in the dataset is primarily driven by the size-related features of the property

- Permutation Feature Importance results (Linear Regression model):

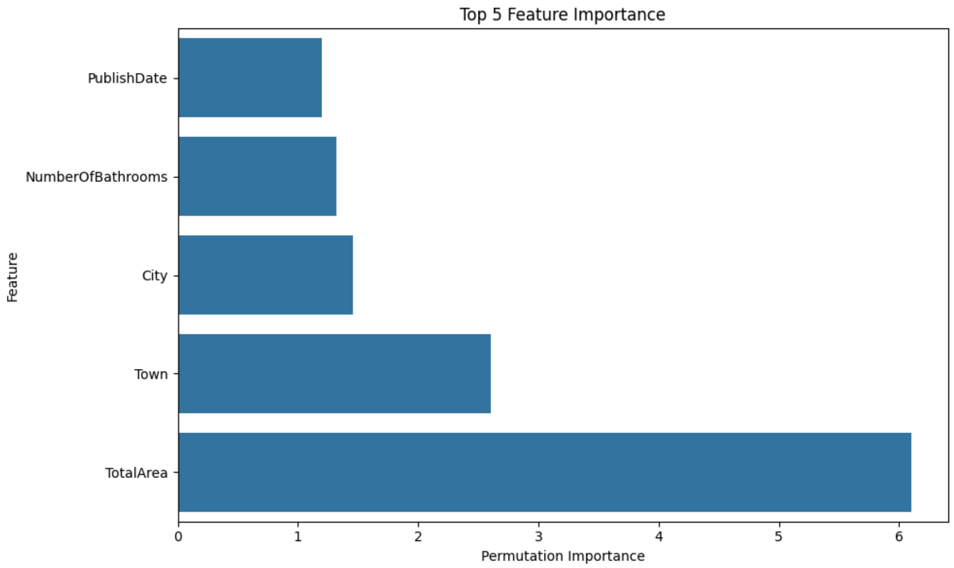


Figure 3: Feature Importance barplot of Real Estate in Portugal using Linear Regression model

Conclusion: It prioritizes properties with larger space in desirable towns and cities

Dataset 2: London housing

- Feature Importance results (Random Forest Regressor model):

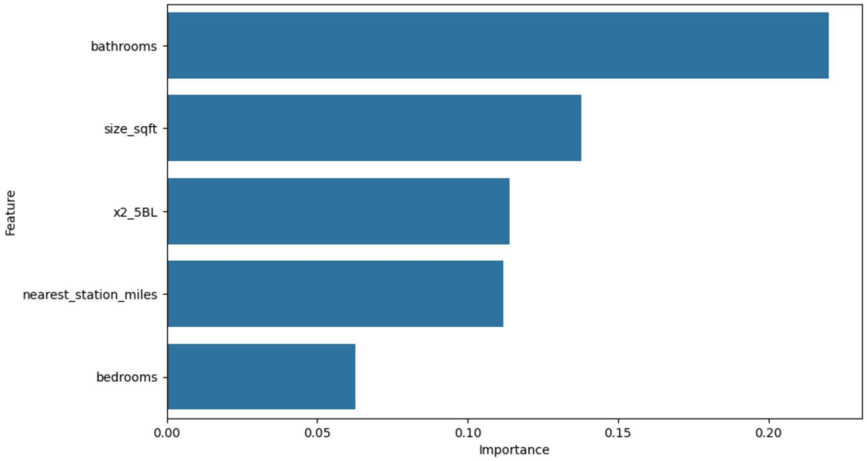


Figure 4: Feature Importance barplot of Real Estate in London using Random Forest Regressor model

Conclusion: Buyers seem to prioritize interior comfort and functionality over external factors

- Permutation Feature Importance results (Linear Regression model)

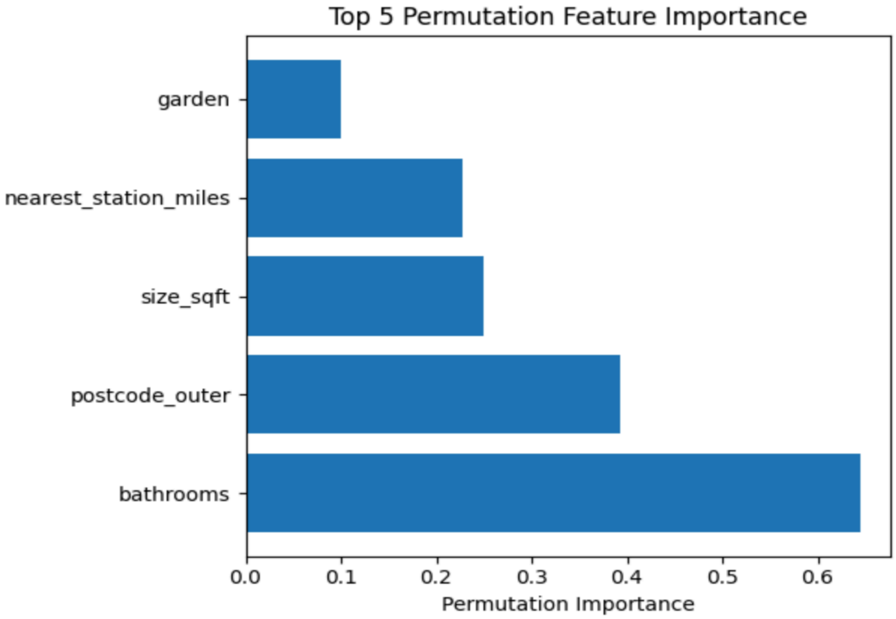


Figure 5: Feature Importance barplot of Real Estate in London using Linear Regression model

Conclusion: prioritize properties with multiple bathrooms in desirable postcode areas, and consider the size and proximity to public transportation.

Overall conclusion on both datasets:

The feature importance results across datasets as well as several prior research consistently highlight size, functionality, and location as key drivers of housing prices.

- Size-Related Features: Built area, total area, living area, and lot size are among the top contributors in both datasets. This justifies their inclusion as primary model inputs, as larger properties consistently command higher prices.
- Interior Comfort & Functionality: Bathrooms and bedrooms significantly impact house prices, especially in London. Buyers prioritize properties with multiple bathrooms, making them an essential model variable.
- Location & Accessibility: Features like town, city, postcode, and proximity to transport influence pricing, although their impact varies by dataset. These inputs are necessary to capture regional price variations and accessibility preferences.
- Additional Influential Factors: Categorical variables such as garden also contribute, indicating that neighborhood characteristics and outdoor space should not be overlooked in price prediction models.

With Warsaw being the capital and economic center of Poland, the real estate market has undergone significant changes in recent years, especially after COVID-19 and welcoming a new wave of foreigners entering Poland.

The first thing to note is that the city has invested heavily in its metro and public transport network, which has recently been converted to trams, partly demonstrating the government's sustainability and concern for the environment. Needless to say, this has had a significant impact on real estate prices in well-connected districts such as Śródmieście, Mokotów and Ochota. The construction of the new M2 metro line is also a big reason for the increase in real estate prices around M2 metro stations, as access to public transport has changed significantly.

Another reason is the development trends of different districts. With the central and older districts attracting residents and young adults due to the employment opportunities provided by major corporate headquarters as well as high-quality amenities and good urban infrastructure. Meanwhile, the suburban districts such as Białołęka and Ursus are undergoing rapid urbanization, making them attractive to families looking for affordable housing options that can be larger than those in the city centre.

What cannot be overlooked is the Sustainability trend from ESG Policies. With an increasing focus on sustainability, Warsaw has been implementing stricter energy efficiency standards. Green building standards (such as BREEAM, LEED) are gaining priority in the real estate market, as they meet the ESG needs of investors and reduce operating costs. Therefore, ESG trends can impact asset values and future project development strategies.

3. Model Design

Considering the nature of residential price prediction, possible modeling techniques can be used are:

- Linear Regression: Simple and interpretable but may not capture complex relationships.
- Random Forest: Handles non-linearity and interactions well, robust to outliers.
- Gradient Boosting Machines (GBM, XGBoost, etc.): Highly accurate but computationally expensive.
- Neural Networks: Suitable for large datasets with complex patterns, but lacks interpretability.

For a short-term forecast horizon (less than 3 months), models like Random Forest or Linear Regression may provide strong predictive power with relatively fast training times. For a long-term forecast horizon (more than 2 years), more complex models like Neural Networks can be considered to capture evolving patterns.

In the project, we propose using the Random Forest Regressor model, which may handle non-linearity relationships better than Linear Regression and be balanced between accuracy and interpretability.

The target variable of the model is the final price of the house (in Polish złoty), which represents the latest sale price of the apartment. This price might be influenced by temporary market fluctuations.

Model performance can be evaluated using metrics such as R-squared (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE); the below table shows the performance threshold of the model, considered the data limitation challenge.

Metrics/Performance threshold	**Acceptable**	**Moderate**	**Poor**
R^2	> 0.85	0.6 - 0.85	< 0.6
RMSE	< 5% of average price	5 - 15% of average price	> 15% of average price
MAE	< 10% of average price	10 - 25% of average price	> 25% of average price

Table 1: Performance threshold

The Random Forest model does not have strict assumptions like the Linear Regression, however, understanding these implicit assumptions can help tuning the model more effectively in practice:

- There is no need to have a linear relationship between the features and the dependent variable. This is an advantage over the traditional linear regression model.
- The model works best when the input variables are actually related to the output variables. If all the input variables are meaningless, the model may just make random predictions. This

assumption can be checked with correlation matrix, testing the relationship between features and the target variable.

- The decision trees in the forest learn from different data samples and are not completely dependent on each other. This is achieved by using the Bootstrap Aggregating (Bagging) method, where each tree is trained on a different subset of the data.
- As the number of trees increases, the overall error will decrease due to the Law of Large Numbers. This can be tested during the optimization process of the model building phase, when we try to increase the number of estimator parameters of the model.
- The model assumes that some variables have more influence on the predicted outcome. This can be tested by feature importance visualization, which shows the effects of features to the target variable.

The method that can be used to measure a model's interpretability is the feature importance score from the Random Forest model, which describes the impurity reduction of each feature.

C. Proof of Concept (PoC)

1. Data Pipeline:

- **Target variable:** final price of an apartment (Polish złoty)
- **Dataset source:** The base dataset, which is created by Dawid Cegielski, contains 10 columns and 23.8k rows, shows the price of apartments around Poland
    - address - Full address
    - city - Warszawa (Warsaw), Kraków (Cracow), Poznań (Poznań)
    - floor - The number of the floor where the apartment is located
    - id - id
    - latitude - latitude
    - longitude - longitude
    - price - Price of apartment in PLN
    - rooms - Number of rooms in the apartment
    - sq - Number of square meters of the apartment
    - year - Year of the building / apartment
- **Data manipulation steps:**
    - Filter to get apartment in Warsaw
    - Extract district from address to create the “district” feature
    - Using longitude and latitude to calculate the distance to the nearest metro station, creating the “distance_to_nearest_metro” feature
    - Using longitude and latitude to calculate the distance to the centre of Warsaw (Palace of Culture and Science), creating the “distance_to_pkin” feature
    - Creating the “living_condition_index”, the calculating method is mentioned below
    - Crawling data from the air quality website, mapping crime_rate by district, creating the es_index, the calculating method is mentioned below
    - Filter years that after 2021 as they seem to be mistake data (the dataset is created in 2021)

- Using years column to create the “age_of_house” feature
- Drop unnecessary columns
- Correct data type of the features

After all of the above steps, we have the final dataset with 8926 rows and 9 features

Currently, the team has included some factors that reflect market trends in the dataset, but has not fully integrated the factors that affect house prices. Variables such as district, nearest_metro_station, living_condition_index, and es_index have partly reflected the changes in the housing market over time. However, the dataset still lacks other important factors such as credit conditions, demographics, and income, which are key factors affecting housing supply and demand.

Not fully integrating market factors can make the house price prediction model inaccurate. If the model only relies on variables about location and living conditions without taking into account changes in housing demand due to population growth, migration trends, or fiscal policy, the prediction may be biased. For example, during periods of easy credit, house prices may be higher because more people can afford to buy houses, while periods of tight credit may cause house prices to fall.

We are able to find these following market-trend related variables:

Public transportation plays an indispensable role in elevating property values through improved accessibility, economic growth, sustainable development, and positive socioeconomic factors. As urban development continues to evolve, the integration of public transportation remains a crucial element in shaping property trends and ensuring lasting value for homeowners and investors alike. Therefore, we have the nearest_metro_station variable.

According to Analysis of Accessibility of Public Transport in Warsaw in the Opinion of Users, research shows that proximity to metro stations increases real estate prices due to the convenience of travel and access to urban amenities. Therefore, more and more people are planning to find housing with the closest distance to public transport stations, including metro, which is the public transport with the fastest average travel time. Public transportation is essential in increasing property values by enhancing accessibility, driving economic growth, promoting sustainable development, and contributing to positive socio economic changes. As cities continue to develop, integrating public transit plays a vital role in shaping real estate trends and maintaining long-term value for both homeowners and investors.

The **district** variable reflects the development of different areas in the city. With a tendency to choose to live in the central districts, people prefer central districts such as Śródmieście, Mokotów, and Ochota due to the high quality of life, full urban facilities, and good job opportunities. These districts have an attractive labor market, attracting people from other areas to live and work. The peripheral districts such as Białołęka, Ursus, and Bemowo are undergoing rapid urbanization, with the development of new residential areas and infrastructure improvements. Housing prices are lower than in the central districts, attracting young people and new families. Therefore, the district variable can reflect the market trend quite well in the scope of house price prediction

With the **Living_condition_index** variable, it reflects the trend of improving living conditions and infrastructure, an important factor influencing the decision to buy a house.

Finally, with **ES_index**, this variable reflects the trend of sustainable development and environmental policy (ESG). More and more people are interested in environmental factors when choosing to buy a house, which may affect the value of real estate in the future.

ESG (Environmental, Social, and Governance) is a set of criteria that assess environmental, social, and governance factors that affect the value and performance of real estate. In this project, we attempted to integrate ESG factors into the dataset; however, we decided to use only the ES Index instead of the ESG Index due to limitations in data collection. Specifically as follows:

- Environmental:

  Air Quality Index (AQI) data is extracted from websites providing environmental information in Warsaw. This variable reflects the level of air pollution, an important factor affecting the quality of life and real estate value.

- Social:

  Crime rate data is collected from Warsaw safety statistics reports. This variable helps assess the safety level of the area, an important factor for home buyers.

- Governance:

  We are facing limitations coming from gathering the Governance factor due to the limited scope of the dataset in Warsaw. Governance data is often related to macro factors such as government policies, urban regulations, or local governance performance, which are difficult to collect and do not fit the scope of the dataset. This reduces the ability to comprehensively assess ESG factors.

  Although the use of the ES Index instead of the ESG Index reduces the comprehensiveness of the index, it still reflects important environmental and social factors. The integration of the ES Index has improved the ability to predict real estate prices based on sustainability criteria. However, the limitation of Governance data needs to be addressed in future studies to achieve a more comprehensive assessment.

- **Model features:**

Feature	Description	Direction of correlation to the target variable
floor	The number of the floor where the apartment is located	Unclear relationship
rooms	Number of rooms in the apartment	Unclear relationship
sq	Number of square meters of the apartment	Positive
district	District where the apartment is located	Central districts will have higher apartment prices

		than suburban districts
distance_to_nearest_metro	Distance from the apartment the nearest metro (km)	Negative
distance_to_pkin	Distance from the apartment to the centre of the city (Palace of Culture and Science) (km)	Negative
age_of_house	Age of the apartment	Negative
living_condition_index	The index shows the attractiveness of living condition around the apartment * The calculating method is shown below	Positive
es_index	The index shows the environment and social around the apartment * The calculating method is shown below	Negative

Table 2: Model inputs

The **living condition index** is taken from the report RANKING OF WARSAW DISTRICTS ACCORDING TO THE ATTRACTIVENESS OF LIVING CONDITIONS from Statistical Office in Warsaw, in which 11 criteria below are aggregated:

x1 – the number of beneficiaries of social assistance per 1 thousand population;

x2 – the share of parks, lawns and green belts in residential estates in the total area (%);

x3 – share of communication function in total area (%);

x4 – number of people per 1 facility of any of the kind: cinemas, theatres, museums (sum);

x5 – number of children in nursery schools per 1 thousand children aged 3-6;

x6 – pupils per 1 personal computer with access to the Internet in primary schools;

x7 – number of people per 1 health care establishment;

x8 – number of people per 1 pharmacy;

x9 – number of people per 1 library;

x10 – number of people per 1 shop;

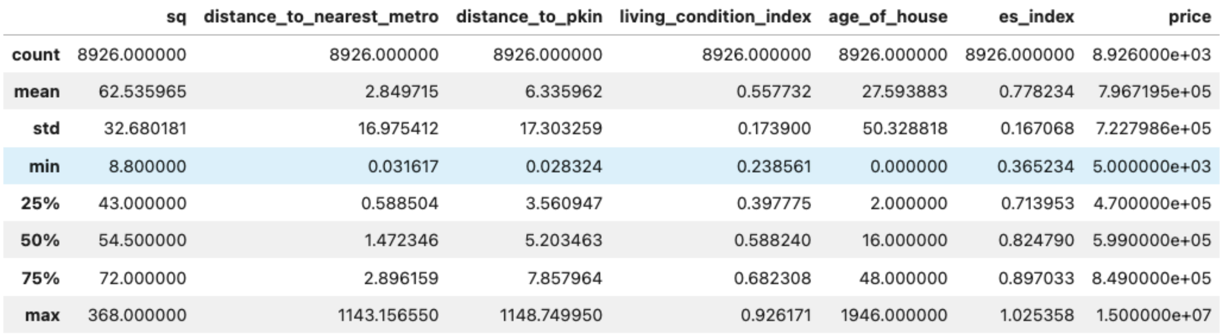
x11 – length of cycling paths (in km);

When putting it into our project dataset, to avoid the apartments in the same district having the same value, which would not contribute much to the apartment price determination, we decided to add a *noise* - range of index by district, and the houses in each district will receive a random living condition index value according to its district. For example: the Ochota district has the living_condition index of 0.673, we create a range with min_index = 0.9*0.673, max_index = 1.1*0.673, and an apartment in the Ochota district received a random value in this range.

The **es_index** is aggregated by 2 metrics, air quality index of the area and the crime rate of the area. This index assumes the weights of above metrics are equal, and they are normalized to the scale of 0 to 1 and subtracted by 1. The data of AQI by district is crawled from the air quality report website, which can be found here and the crime rate is calculated by dividing the number of crimes by the population of each district in Warsaw, the detailed report of the number of crimes by district in Warsaw can be found here. Similar to the living_condition index, the es_index is also added with noise - range of index by district.

These two indexes, combined with some features such as “distance_to_nearest_metro”, “distance_to_pkin” reflect somehow the trend of the real estate market in Warsaw as well as the effects of the environment - social - governance to the price of apartments.

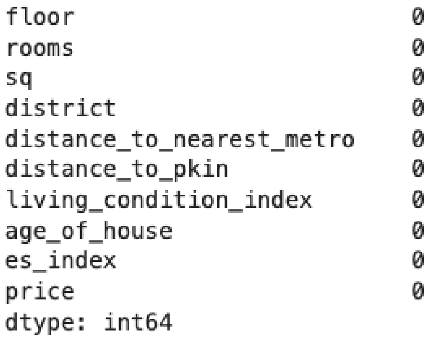
- **Summary statistics of features in the final dataset:**
  - **Basic Statistics:**



	sq	distance_to_nearest_metro	distance_to_pkin	living_condition_index	age_of_house	es_index	price
count	8926.000000	8926.000000	8926.000000	8926.000000	8926.000000	8926.000000	8.926000e+03
mean	62.535965	2.849715	6.335962	0.557732	27.593883	0.778234	7.967195e+05
std	32.680181	16.975412	17.303259	0.173900	50.328818	0.167068	7.227986e+05
min	8.800000	0.031617	0.028324	0.238561	0.000000	0.365234	5.000000e+03
25%	43.000000	0.588504	3.560947	0.397775	2.000000	0.713953	4.700000e+05
50%	54.500000	1.472346	5.203463	0.588240	16.000000	0.824790	5.990000e+05
75%	72.000000	2.896159	7.857964	0.682308	48.000000	0.897033	8.490000e+05
max	368.000000	1143.156550	1148.749950	0.926171	1946.000000	1.025358	1.500000e+07

Table 3: Summary statistics of variables

- **Missing Data:** there are no missing values in the dataset.



```
floor                        0
rooms                        0
sq                           0
district                     0
distance_to_nearest_metro    0
distance_to_pkin             0
living_condition_index       0
age_of_house                 0
es_index                     0
price                        0
dtype: int64
```

- **Data Distribution**
  - **Categorical Variables:**

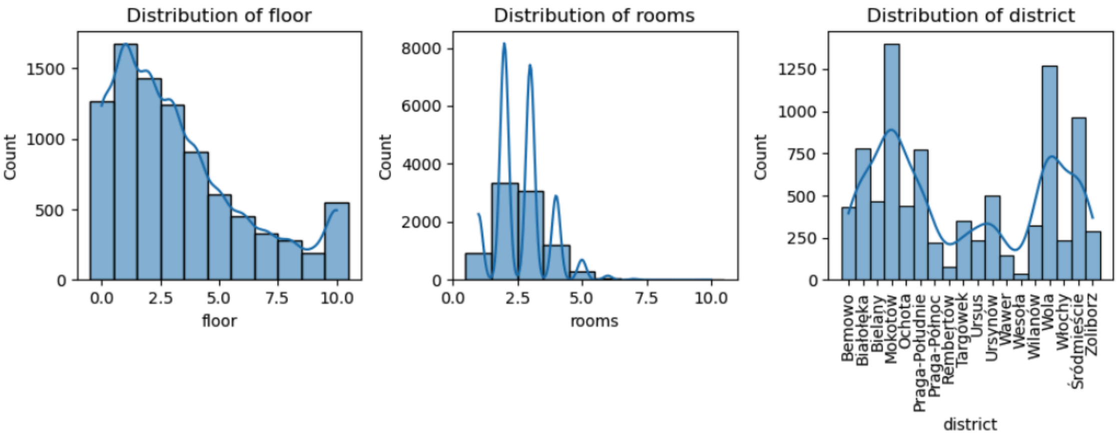


Figure 6: Categorical features distribution

Observation:

- floor and rooms are also right-skewed, with many values concentrated on the left side and some extremely large values on the right side.

Suggested action: Right-skewed variables can degrade model performance, especially models that rely on normal distribution assumptions such as linear regression. To improve model performance, we need to apply data transformation methods to reduce skew.

  - **Numeric Variables:**

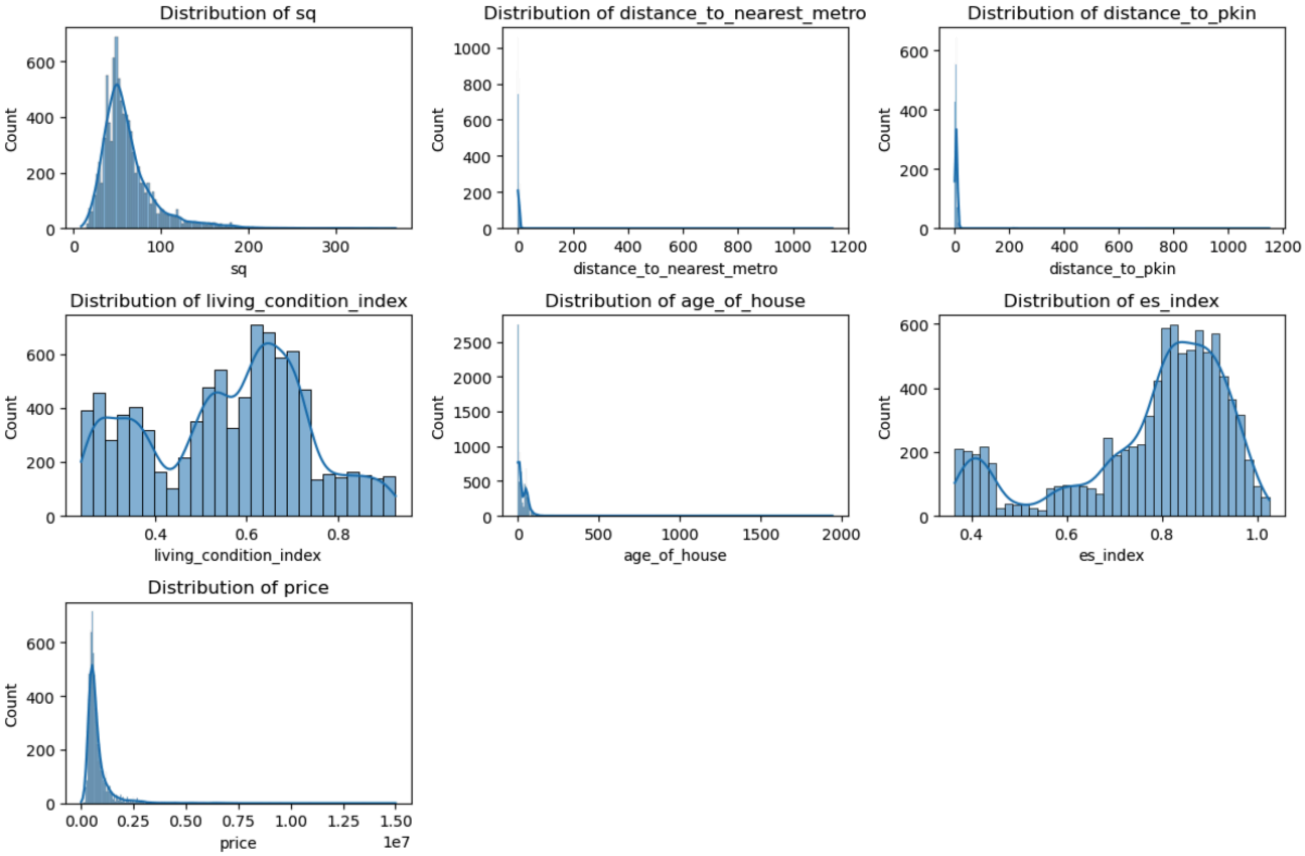


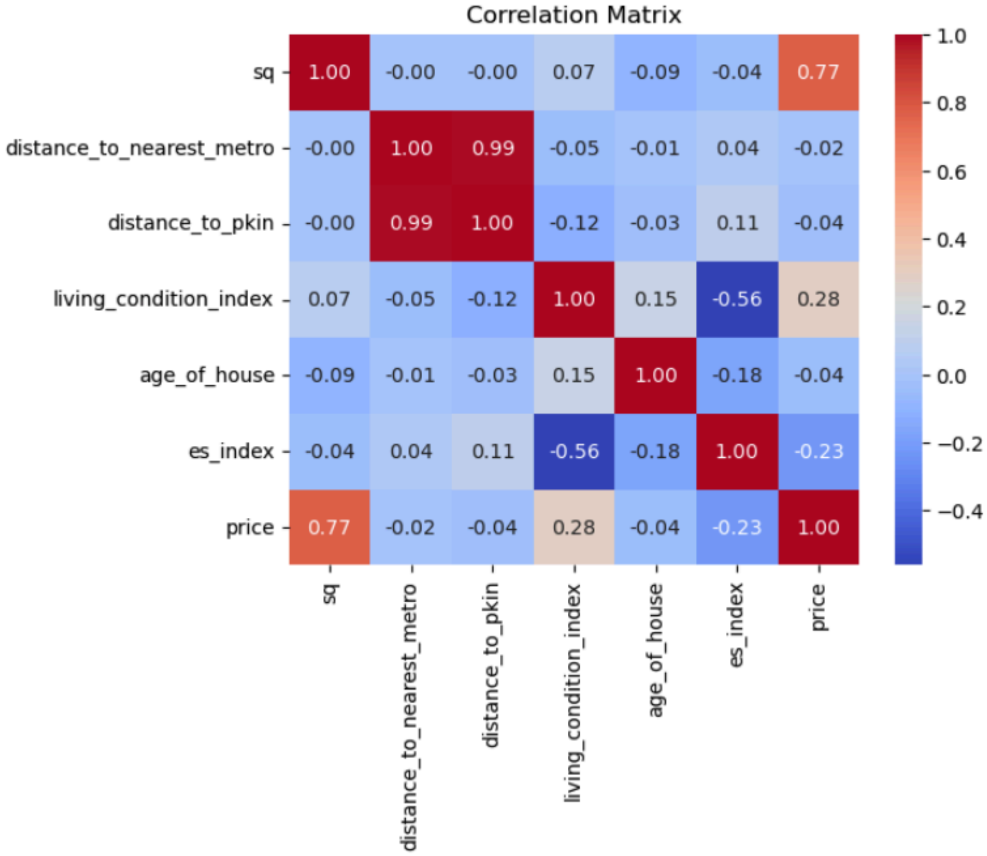
Figure 7: Numerical features distribution

Observation:

- Features except living_condition_index and es_index are all right-skewed
- We need to deal with long tail variables (outliers) like sq, distance_nearest_metro, age_of_house or distance_to_pkin
- Moreover, the target variable - price is also right-skewed, but outliers are valid values (e.g. cheap houses in a particular area), removing them may lose important information.

Suggested action: we can standard scale or use transformation methods later in model training.

- **Correlation matrix**



	sq	distance_to_nearest_metro	distance_to_pkin	living_condition_index	age_of_house	es_index	price
sq	1.00	-0.00	-0.00	0.07	-0.09	-0.04	0.77
distance_to_nearest_metro	-0.00	1.00	0.99	-0.05	-0.01	0.04	-0.02
distance_to_pkin	-0.00	0.99	1.00	-0.12	-0.03	0.11	-0.04
living_condition_index	0.07	-0.05	-0.12	1.00	0.15	-0.56	0.28
age_of_house	-0.09	-0.01	-0.03	0.15	1.00	-0.18	-0.04
es_index	-0.04	0.04	0.11	-0.56	-0.18	1.00	-0.23
price	0.77	-0.02	-0.04	0.28	-0.04	-0.23	1.00

Figure 8: Correlation matrix of the variables

Observation:

- The most valuable variables for model prediction: sq with strong correlation (0.77), followed by living_condition_index when it has a positive correlation with price.
- distance_to_nearest_metro and distance_to_pkin have very strong correlation with each other (0.99).

2. Model Output:

The Random Forest Regressor model was trained using the selected dataset, ensuring that the data preprocessing steps, including standardization, encoding categorical features, were properly conducted.

The assumption that the selected features contribute meaningfully to the target variable is performed by the feature importance analysis, using the mean decrease in impurity (MDI) as the key driver. The findings were compared with prior research and literature, confirming that the selected features align with existing economic and financial studies.

Another assumption of the model, which stated the larger number of trees lead to the reduction of error, is also tested during the model training process.

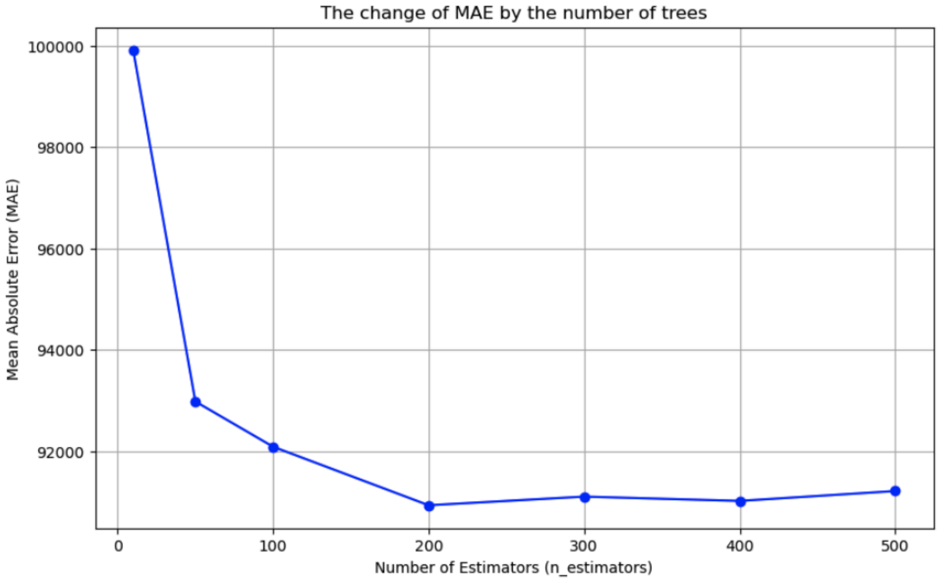


Figure 9: Mean absolute error by number of trees

The graph shows that as the number of trees in the model increases, the MAE index decreases, however, at a certain threshold, the error decreases slowly and even becomes a bit higher.

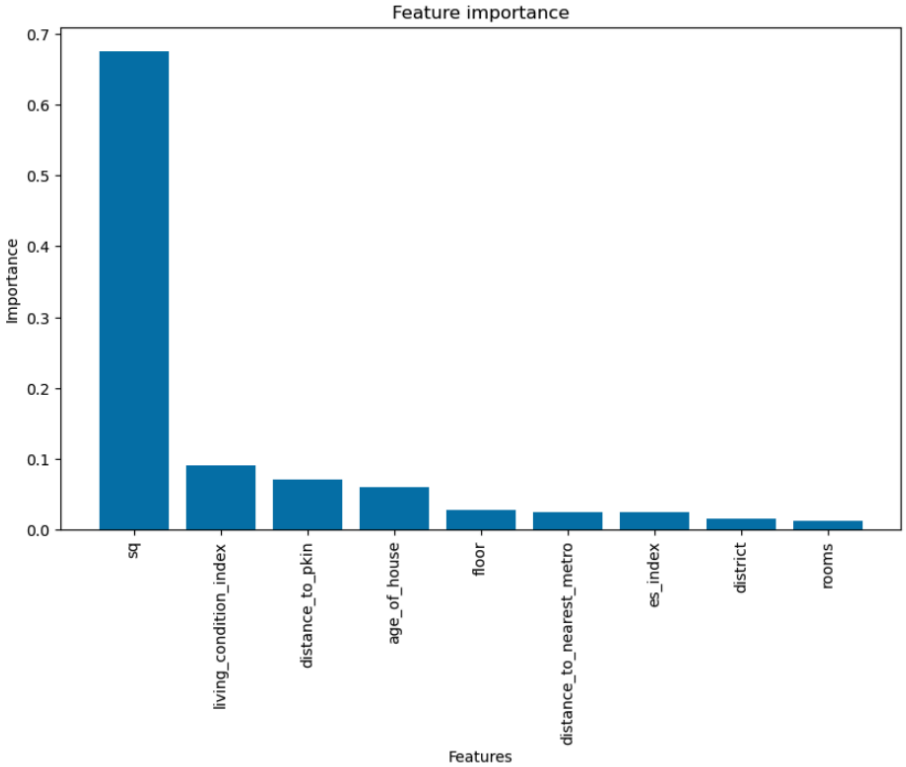


Figure 10: Feature Importance of the model

The above chart shows the most important features contributing to the price of an apartment, ensuring business end-users can understand the model's decision. According to the chart, we can see that the acreage factor contributes largest to the price of an apartment in the random forest regressor model, followed by some spatial factors such as living conditions and distance to the centre of the city.

The ESG factor, which is included using the es_index feature, also contributes to the decision system of the model. However, it should be emphasized that this feature is artificially created, so there is a risk that it may be unrelated to the reality.

The model's performance was evaluated using the chosen target metric, which includes:

- **R² Score: 0.899**
- **Mean Absolute Error (MAE): 91501 (11.48% of the average price)**
- **Root Mean Squared Error (RMSE): 205764 (25.82% of the average price)**

The R² Score of 0.899 shows the model explains almost 90% percent of the variation of the price of the apartment, while the MAE and RMSE values show model error in predicting the apartment's price. According to the performance threshold, the R-Squared score is great, the MAE is moderate and RMSE is poor.

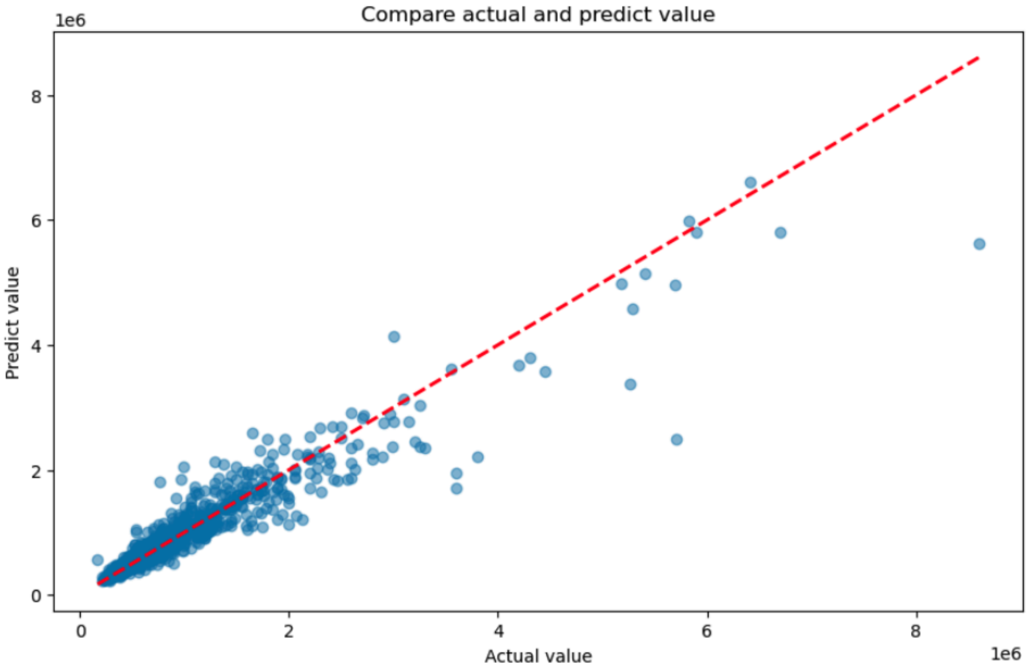


Figure 11: Actual and Predict value plot

The actual and prediction value from the model plot shows that the model predictions have bigger errors for high-priced apartments, which may be due to the lack of data on expensive apartments in the training dataset, causing the model to tend to predict lower apartment prices than their actual value. It can be seen that the model still works relatively well for mid- to low-priced apartments.

Based on these insights, future enhancements may include:

- Using a larger dataset, ensuring an even distribution of house prices from low to high
- Integrating gradient boosting models or experimenting with neural network approaches to capture non-linear patterns more effectively although this may lead to the trade-off of model interpretability

Model's complexity:

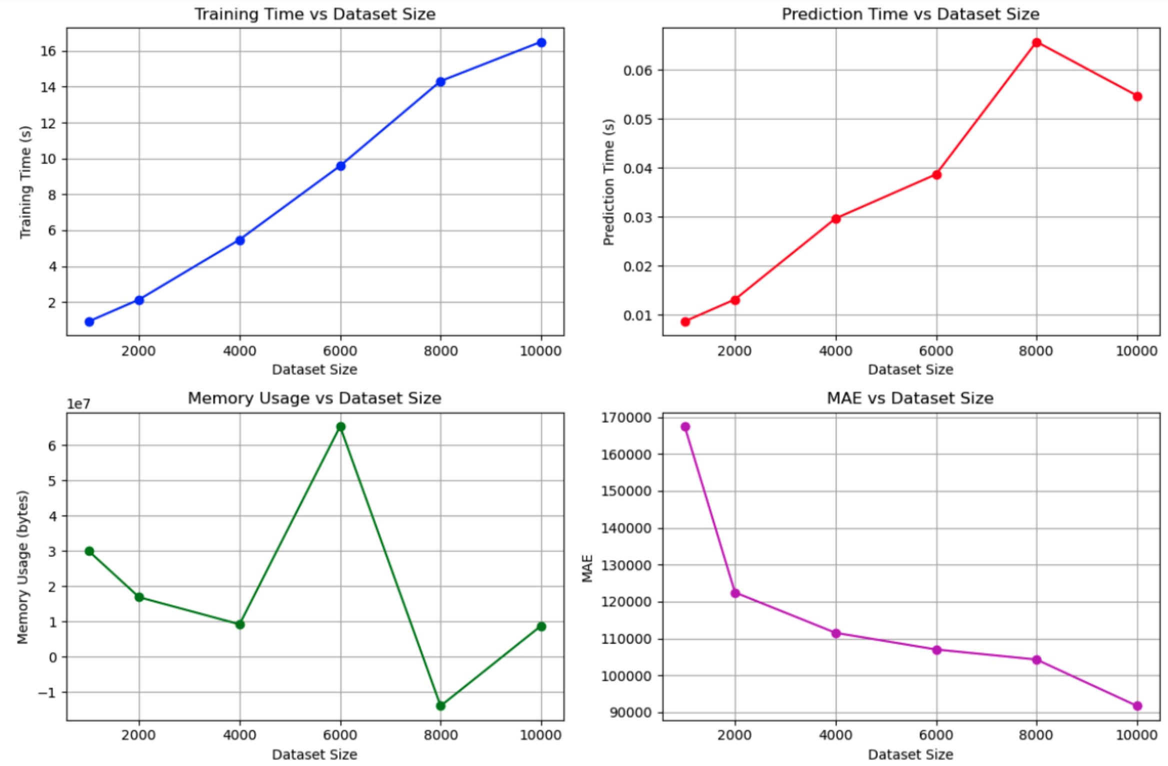


Figure 12: Model's time and memory usage by dataset size

Above charts show the model's time and memory usage by the dataset size. The time for the model to learn is increasing when the dataset size is increasing, the same trend is also seen in the time for the model to make predictions. The MAE line is also decreasing along with the increasing of the dataset size, showing that if we have more data, the model's accuracy may be improved.

D. GenAI Usage

GenAI has contributed greatly to the project in facilitating many tasks and research processes. From the beginning, the tool has helped in structuring the report outline and concluded the key components of the project. In this way, time is used effectively and the delegation of the tasks has been conducted with ease. GenAI has been put into practice for aggregating the dataset by combining different sources into a final dataset and optimizing the process of cleaning the data dynamically. Beyond the optimization usage, the tool has further proposed many creative solutions to driving questions throughout the project including the method of calculating the living condition index based on several factors affecting the quality of life and ES_index reflecting the sustainable development and environmental policy. We can clearly see the potential implications of the tool. Nonetheless, we have not been purely dependent on the output of GenAI, but consciously use the conventional ways in evaluating the project including traditional resources and also the prior knowledge of building the model. The entire team has verified for every detail of the project to ensure the reliability and accuracy of the findings.

E. Discussion and Conclusion

Achievements

In this project, we successfully developed a residential price prediction model using Random Forest Regressor. The model evaluation results are as follows:

- R^2 Score: 0.899, indicating that the model explains approximately 89.9% of the variance in housing prices.
- Mean Absolute Error (MAE): 91,501 (11.48% of the average price), showing a reasonable level of absolute prediction accuracy.
- Root Mean Squared Error (RMSE): 205,764 (25.82% of the average price), reflecting the average error magnitude.

The Random Forest Regressor was chosen due to its ability to handle non-linearity, manage missing values, and provide feature importance insights. Additionally, by leveraging feature important values, we ensured that the model's decision-making process remains interpretable, enabling end users to understand which factors drive housing prices.

Challenges and Limitations

Despite the promising results, several challenges and limitations were encountered:

- Dataset Size and Composition: The dataset used in this project is not particularly large. The core dataset primarily contains property-specific features, while additional factors such as ESG (Environmental, Social, and Governance) metrics, spatial data, and external market conditions were artificially integrated. While these added variables may enhance predictive accuracy, they introduce the risk of reduced real-world validity, as their representation might not fully capture actual market dynamics.
- Scope Restriction to Warsaw: By narrowing the project's scope to residential price prediction within Warsaw, the process of collecting relevant external data, particularly ESG factors and market trends of the area, became significantly more challenging. The availability of reliable, granular data for these external influences was limited, which may have affected the model's ability to generalize to broader housing markets.
- Model Sensitivity to Feature Types: The model primarily relies on structured numerical and categorical features, making it less effective in capturing qualitative aspects that play a crucial role in real estate valuation. Key elements such as neighborhood prestige, architectural design, and future urban development plans are not explicitly included, which could lead to gaps in the

model’s predictive capability.

Addressing these limitations in future iterations of the project—such as increasing dataset size, incorporating more robust real-world external data sources, and developing methods to integrate qualitative factors—could further enhance the model’s accuracy and applicability.

Possible future improvements

- Explore Hybrid Modeling Approaches: Combining Random Forest with other models (e.g., Gradient Boosting, or Linear Regression for residual corrections) may lead to better generalization.
- Experiment with Deep Learning: Using complex model like Neural Networks may help as it could learn the non-linear relationship of the variables.
- Enhance Model Interpretability: Besides Feature Importance values, additional techniques like SHAP, LIME (Local Interpretable Model-Agnostic Explanations) and Partial Dependence Plots (PDP) can provide further insights into individual predictions and feature interactions.

F. References

1. Warsaw City Council (2024) *Warsaw: Your place to invest 2024*. Available at: https://en.um.warszawa.pl/documents/12024633/17756640/Warsaw+Your+place+to+invest+2024.pdf/035f36b2-76e6-931d-e9a9-1de2479bbcd5?t=172854766214
2. JLL (2021) *Warsaw City Report Q2 2021*. Available at: https://www.jll.pl/content/dam/jll-com/documents/pdf/research/emea/poland/en/jll-pl-en-warsaw_city_report_Q2_2021_2.pdf
3. Investopedia (n.d.) *Hedonic Pricing*. Available at: https://www.investopedia.com/terms/h/hedonicpricing.asp (Accessed: [Ngày truy cập]).
4. Price Creek Event Center (n.d.) *The Role of Public Transportation in Increasing Property Values*. Available at: https://pricecreekeventcenter.com/the-role-of-public-transportation-in-increasing-property-values/#:~:text=Public%20transportation%20plays%20a%20critical,property%20demand%20and%20increased%20values
5. Warsaw Statistical Office (n.d.) *Ranking of Warsaw Districts in Terms of Attractiveness of Living Conditions*. Available at: https://warszawa.stat.gov.pl/files/gfx/warszawa/en/defaultaktualnosci/810/2/1/1/ranking_dzielnic_warszawy_pod_wzgledem_atrakcyjnosci_warunkow_zycia_ang.pdf
6. ResearchGate (2023) *Analysis of Accessibility of Public Transport in Warsaw in the Opinion of Users*. Available at: https://www.researchgate.net/publication/367866688_Analysis_of_Accessibility_of_Public_Transport_in_Warsaw_in_the_Opinion_of_Users
7. National Bank of Poland (2021) *Report on the Situation in the Residential and Commercial Real Estate Market in Poland in 2021*. Available at: https://nbp.pl/wp-content/uploads/2025/01/Report-on-the-situation-in-the-residential-and-commercial-real-estate-market-in-Poland-in-2021.pdf
8. CBRE (2021) *Poland Market Outlook 2021*. Available at: https://assets.cavatina.pl/downloads/projects/Poland_Market_Outlook_2021_CBRE.pdf
9. Warsaw Stock Exchange (n.d.) *ESG Reporting Guidelines*. Available at: https://www.gpw.pl/pub/GPW/ESG/ESG_Reporting_Guidelines.pdf
10. Luvathoms (2024) Portugal Real Estate 2024. Available at: https://www.kaggle.com/datasets/luvathoms/portugal-real-estate-2024
11. Bertie Mackie (n.d.) London House Listings. Available at: https://www.kaggle.com/datasets/bertiemackie/london-house-listings