Faculty of
Mathematics
and Informatics

VILNIUS UNIVERSITY
FACULTY OF MATHEMATICS AND INFORMATICS
DATA SCIENCE
MASTER'S STUDY PROGRAMME

PREDICTING REAL ESTATE PRICES WITH MACHINE LEARNING

Master's thesis

Author: Blessing Oritsewenyimi Agadagba
VU email address: blessing.agadagba@mif.stud.vu.lt
Supervisor: (Dr. Tomas Plankis)

Vilnius

2024

Santrauka

Šiame magistro darbe nagrinėjami įvairūs mašininio mokymosi algoritmai, tokie kaip tiesinė regresija, Ridge, Lasso, ElasticNet, sprendimų medis, atsitiktinių miškų modelis, XGBoost, LightGBM ir gradiento didinimas, siekiant prognozuoti nekilnojamojo turto kainas, naudojant Vilniaus savivaldybės duomenis. Tyrimas apima duomenų valymo ir paruošimo, modeliavimo bei hiperparametrų optimizavimo etapus. Tyrimas vykdomas sistemingai, pradedant baziniu modeliu, kuris naudoja tiesinę regresiją tik su butų duomenimis. Siekiant įvertinti papildomų demografinių kintamųjų ir naujai sukurtų požymių įtaką modelio veikimui, jie buvo įtraukti į analizę. Modeliai buvo optimizuoti ir vertinami, remiantis trimis duomenų rinkinio konfiguracijomis: (1) tik butų duomenys, (2) butų duomenys kartu su demografiniais kintamaisiais ir (3) butų duomenys, papildyti tiek demografiniais kintamaisiais, tiek naujai sukurtais požymiais. Rezultatai parodė, kad ansambliniai ir didinimo modeliai žymiai pranoko tiesinius modelius. LightGBM pasiekė didžiausią prognozavimo tikslumą (R^2 = 0.8698, RMSE = 35,714, MAPE = 12,54 %), kai buvo naudojami butų duomenys kartu su demografiniais ir naujai sukurtais požymiais.

Raktiniai žodžiai: Nekilnojamojo turto kainų prognozavimas, LightGBM, Ansamblio mokymasis, XGBoost, Linijinė regresija, Savybių inžinerija, Mašininis mokymasis, Vilniaus savivaldybė

Abstract

This master's thesis explores a range of machine learning algorithms, including linear regression, Ridge, Lasso, ElasticNet, Decision Tree, Random Forest, XGBoost, LightGBM, and Gradient Boosting, for predicting real estate prices using data from Vilnius Municipality.

The study outlines the steps for data cleaning and preparation, modelling, and hyperparameter tuning. The research follows a systematic approach, beginning with a baseline model using linear regression on apartment data alone. Additional demographic variables and newly engineered features were incorporated to assess their impact on model performance.

The models were tuned and evaluated using three distinct configurations of the dataset: (1) apartment data alone, (2) apartment data combined with demographic variables, and (3) apartment data augmented with both demographic variables and newly engineered features. The results reveal that ensemble and boosting models significantly outperformed linear models, with LightGBM achieving the highest predictive accuracy ($R^2 = 0.8698$, RMSE = 35,714, MAPE = 12.54%) when apartment data with both demographic and newly engineered features were utilised.

Keywords: Real Estate Price Prediction, LightGBM, Ensemble Learning, XGBoost, Linear Regression, Feature Engineering, Machine Learning, Vilnius Municipality

List of Figures

1 Overall distribution of real estate listings by neighborhood. 28
2 Distribution of real estate listings by neighborhood after removing districts in the outskirts. 29
3 Q-Q Plot of Price Variable. 30
4 Histogram of property prices with percentile thresholds. 30
5 Distribution of property prices. 32
6 Average Price distribution based on number of floors 33
7 Construction year over Property price 34
8 Average property prices distribution per Nerighbourhood 34
9 Maximum and Mininum property Prices per Neighbourhood 35
10 Distribution of Building type across apartment listing 35
11 Building Type Distribution by Neighbourhood 36
12 Scatter plot showing the relationship between property area and Property price 37
13 Scatter plots showing the relationship between property price (in Euros) and proximity (in meters) to key amenities 38
14 Scatter plots showing the relationship between demographic features (total number of children, single individuals, and total population) and property prices 39
15 Correlation heatmap of numerical features 40
16 Baseline Model: Actual vs. Predicted Values for Linear Regression. 46
17 Comparison of Predicted vs. Actual Values for Baseline and Best Models 50
18 Top 20 LightGBM Feature Importance 52
19 Distribution of building type by neighbourhood with percentages 61

List of Tables

1 Summary of Results 24
2 Hyperparameter Grids for Model Tuning 42
3 Baseline Linear Regression — Test Results (Group 1) 46
4 Model Performance (Apartment Data only) 47
5 Model Performance (Apartment Data + Demographics Data) 48
6 Model Performance (Apartment + Demographics + Newly Engineered Features) 49
7 Best Models Across Different Data Groups 50
8 Group 1 Features: Apartment Data only 62
9 Group 2 Features: Apartment + Demographics 62
10 Group 3 Features: Apartment + Demographics + ewly Engineered Features 62
11 Engineered Features and Their Calculations 64
12 Best Hyperparameters for Models Across Data Groups 66

Contents

1 Introduction **7**
- 1.1 Background of study . . . 7
- 1.2 Research Problem . . . 8
- 1.3 Research Rationale . . . 8
- 1.4 Research Aim and Objectives . . . 8
- 1.5 Research Questions . . . 9
- 1.6 Research Significance . . . 9
- 1.7 Research Methodology . . . 10
- 1.8 Report Organisation . . . 10

2 Literature Review **11**
- 2.1 Factors Influencing Real Estate Prices . . . 11
- 2.2 Traditional Real Estate Price Prediction Methods . . . 13
  - 2.2.1 Hedonic Pricing Models . . . 13
  - 2.2.2 Linear Regression Models . . . 14
  - 2.2.3 ARIMA Models . . . 17
- 2.3 Machine Learning Techniques for Real Estate Price Prediction . . . 18
  - 2.3.1 Decision Tree . . . 18
  - 2.3.2 Random Forests . . . 18
  - 2.3.3 Gradient Boosting Machines (GBM) . . . 19
- 2.4 Ensemble Methods in Real Estate Price Prediction . . . 21
  - 2.4.1 Bagging (Bootstrap Aggregating) . . . 21
  - 2.4.2 Boosting . . . 21
  - 2.4.3 Stacking . . . 21
- 2.5 Related Work . . . 21
- 2.6 Comparative Analysis of Machine Learning Techniques in Real Estate Price Prediction . . . 25
- 2.7 Conclusion . . . 25

3 Methodology **26**
- 3.1 Data Collection . . . 26
  - 3.1.1 Web Scraping Procedure . . . 26
  - 3.1.2 Demographics Data . . . 26
  - 3.1.3 Ethical Consideration . . . 27
- 3.2 Data Cleaning and Preprocessing . . . 27
  - 3.2.1 Unit Standardization . . . 27
  - 3.2.2 Handling Missing Data . . . 27
  - 3.2.3 Removing Districts in the Outskirts of Vilnius . . . 28
  - 3.2.4 Dealing with Outliers . . . 29
- 3.3 Encoding Categorical Variables . . . 31
- 3.4 Feature Scaling . . . 31
- 3.5 Exploratory Data Analysis . . . 32
- 3.6 Feature Engineering . . . 40
- 3.7 Model Selection . . . 41
- 3.8 Hyperparameter Tuning . . . 41
- 3.9 Model Evaluation . . . 42
- 3.10 Conclusion . . . 44

4 Modelling and Results **45**
4.1 Data Splitting . . . 45
4.2 Baseline Model . . . 45
4.3 Results by Data Group . . . 47
4.4 Group 1: Apartment Data Only . . . 47
4.5 Group 2: Apartment Data + Demographics Data . . . 48
4.6 Group 3: Apartment + Demographics + Newly Engineered Features . . . 49
4.7 Comparative Analysis and Discussion . . . 50
4.8 Impact of Demographics . . . 51
4.9 Effectiveness of Feature Engineering . . . 51
4.10 Conclusion . . . 52

5 Conclusion **53**

Appendix 1. Table showing distribution of building type by neighbourhood with percentages **61**

Appendix 2. Feature Groups **62**

Appendix 3. Newly Engineered features **63**

Appendix 4. Model Hyperparameters **65**

Appendix 5. The use of AI tools **67**

Appendix 6. Code Repository **68**

1 Introduction

This chapter introduces the research by discussing the background of the study, outlining the research problem, defining the objectives, setting the questions this research seeks to answer, and describing the study's significance and the research's structure. It provides the foundation for understanding the subsequent chapters and the research as a whole.

1.1 Background of study

Property valuation plays an important role in the economy and is greatly important to a broad spectrum of stakeholders, from individual owners, lenders, land developers to government planning authorities[22]. Lenders and financial institutions must ensure that the value of a property used as collateral aligns with the loan amount [2]. The value of property influences the loan-to-value ratio, which invariably impacts the amount of credit extended to the borrower[2]. Similarly, understanding the market value of a property is important for owners as it helps understand their asset's market value, which is essential for transactions, refinancing, and investment planning. Also, accurate valuations allow owners to better plan tax liabilities and potential return on investments [4]. Like every other stakeholder in the real estate market, the role of property valuation for governments cannot be overemphasised; governments can generate adequate revenue from property taxes, which is essential for funding public services and infrastructure. Discrepancies in property value can lead to unfair taxation. [49]

Within the context of this research, the focus is on the Vilnius municipality, the capital city of Lithuania. Like many European capitals, Vilnius has experienced significant growth and transformation in its real estate market in recent years [50]. This growth has presented both opportunities and challenges [50], making the study of real estate price valuation timely and highly relevant. Several factors, such as economic trends, demographics, governmental policies, and local demand and supply, heavily influence the dynamics of the real estate market [24]. Comprehension of the dynamics of real estate pricing is crucial for individuals and entities aiming to invest wisely and for policymakers responsible for ensuring housing affordability and market stability [35].

Real estate stakeholders and researchers have used various methods for real estate valuation, and the second chapter of this research will review these methods extensively. Traditional models that predict house prices seek to determine the relationship between provided data and house prices. Machine learning in real estate valuation presents significant advantages over traditional models in its ability to process large datasets and manage missing values, identify intricate trends and factors influencing housing prices, and predict house prices more accurately[75]. Leveraging these advanced machine learning techniques can enhance the understanding of the Vilnius real estate market and improve the decision-making process for both buyers and sellers.

This study aims to analyse the factors influencing real estate prices comprehensively to build a model that can accurately predict real estate prices in Vilnius by comparing advanced machine learning techniques.

1.2 Research Problem

Real estate markets are dynamic and complex, posing considerable challenges for homeowners, investors, and policymakers alike. Significant economic growth, urban development, and demographic changes amplify this challenge within the Vilnius municipality[50]. Despite the studies on understanding real estate pricing dynamics, there is a significant gap in knowledge regarding an accurate and efficient predictive model tailored to the unique context of Vilnius. While earlier research has investigated real estate prediction in different global contexts, and research like Grybauskas, Pilinkiene, and Stundziene (2021)[28] explored predicting the revision of real estate prices during the COVID-19 pandemic highlighting the attributes of an apartment that are most likely to influence a price revision during the pandemic [28], there is a lack of models for price prediction that explores the use of advanced machine learning techniques tailored to the current market conditions in Vilnius. This research primarily seeks to develop a reliable model for predicting Vilnius property prices using advanced machine learning.

Investigating factors such as property characteristics, neighbourhood attributes, proximity to amenities, and historical trends that influence real estate prices necessitates an advanced approach capable of accurately capturing and generalising these factors [38]. This research will focus on accounting for these factors in the model development by employing machine learning algorithms and ensuring that the model can adapt to the unique dynamics of the Vilnius market and the availability of an open-source model for real estate price prediction for Vilnius.

1.3 Research Rationale

The rationale for this research lies in the importance of the real estate market as a driving force in urban development and economic stability. Real estate pricing is crucial to the operation of urban economies, affecting various stakeholders, including homebuyers and sellers, property developers and investors, policymakers and urban planners, and the general public. The solution to the research problem has broad implications and benefits. Homebuyers and sellers can make better-informed decisions, resulting in more equitable transactions and reduced risk of making poor investments or accepting subpar offers. For example, real estate professionals, including agents and appraisers, may benefit from a more transparent market with readily available pricing insights, which can increase their capacity to serve customers effectively. Policymakers can employ accurate and realistic pricing models to establish policies that promote housing affordability and sustainable urban development, which could lead to more equitable housing opportunities and reduced socioeconomic disparities. A well-regulated real estate market can maintain overall economic stability by reducing the likelihood of speculative bubbles or crashes. The study's findings extend beyond Vilnius, a significant reference point for other cities with similar real estate pricing challenges. Furthermore, machine learning and data analysis in this context offer valuable insights into the intersection of urban development and technology, with broad implications for both the real estate and data science industries.

1.4 Research Aim and Objectives

This research aims to develop a reliable predictive model for estimating real estate prices in the Vilnius municipality using a comparative analysis of various machine learning algorithms. The study

aims to achieve its goals through the following objectives:

1. Identify machine learning techniques commonly applied to real estate data based on previous studies.
2. Implement machine learning models using linear and ensemble methods.
3. Optimize model performance using hyperparameter tuning techniques like Grid Search.
4. Compare the performance of linear models with ensemble methods.
5. Evaluate models using metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Coefficient of Determination (R^2).
6. Use cross-validation to ensure the models are robust and reliable.
7. Identify the best-performing model and discuss its practical application for real estate price prediction.

1.5 Research Questions

This study will address the following questions:

1. What methods have been employed in previous studies to estimate real estate prices?
2. What are the significant attributes of properties in Vilnius that impact house prices most?
3. Can newly engineered features improve the model's accuracy, and if so, what are these features?
4. Which machine learning algorithm performs best in predicting property prices in Vilnius?

1.6 Research Significance

This study will contribute to the real estate industry and machine learning research. Many industries have widely used machine learning techniques to support data-driven decision-making. However, their application in ensemble techniques to real estate pricing in Vilnius is still relatively novel. By combining advanced algorithms with comprehensive data on property prices and neighbourhood characteristics, this research provides a methodological framework that can adapted for use in different real estate markets. Furthermore, understanding the factors influencing Vilnius real estate prices can have practical implications for housing affordability. Ensuring the affordability of house prices aligns with the societal goals of minimising disparities in socioeconomic status and promoting inclusive urban development. While this study focuses on Vilnius, its findings are not geographically limited. The methods and insights gained directly extend to other urban areas facing similar real estate price challenges.

1.7 Research Methodology

The study will review relevant research related to real estate valuation. Comprehensive data on house prices will be collected from various reputable real estate websites using web scraping techniques. The data will be cleaned, pre-processed, and divided into training and testing sets. Key features will be engineered from the raw data to enhance the model's predictive performance. A range of machine learning algorithms will be explored, including ensemble techniques such as bagging, stacking, and boosting, to identify the most effective model for price prediction. The models will be validated using cross-validation techniques, and their performance will be assessed using root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination (R^2).

1.8 Report Organisation

The thesis is structured as follows: Chapter 1 introduces the study's research problem, objectives, and significance. Chapter 2 provides a comprehensive literature review on real estate price prediction methods and the application of machine learning techniques. Chapter 3 outlines the research methodology, including data collection, pre-processing, and model evaluation. Chapter 4 presents the research findings, including model performance analysis and a discussion of the results. Finally, Chapter 5 concludes the research by summarising the key insights and limitations of the study, discussing the practical implications, and offering recommendations for future research.

2 Literature Review

Predicting real estate prices has long been a focus of researchers and industry professionals due to its complexity and significant economic implications. With advancements in machine learning, various predictive models have emerged, each addressing different aspects of real estate dynamics. This chapter reviews the existing literature on real estate price prediction methods, focusing on applying traditional statistical models and modern machine learning techniques, particularly ensemble approaches. This review provides a comprehensive foundation for the research presented in subsequent chapters.

2.1 Factors Influencing Real Estate Prices

Several factors, varying by location, influence the real estate market[16]. Understanding these factors is essential for developing a reliable prediction model for estimating real estate prices. Economic factors are pivotal in shaping the dynamics of the real estate market[36]. One of the key influencers that shape the real estate market in Lithuania, specifically in Vilnius, is the set of macroeconomic indicators, including factors like GDP growth, inflation rates, interest rates and availability of funding (e.g. availability of bank credits/mortgages) [23]. Improved economic performance boosts overall demand and prices of real estate, housing inclusive [57]. Inflation drives up housing prices and raises the overall prices for goods and services, while deflation causes housing prices to decline.[52]. Rising interest rates increase financing costs, deter potential buyers from purchasing housing, limit real estate market liquidity, and prolong the sales time [53]. As a result, rising interest rates gradually pull down house prices as rent becomes a more appealing alternative [13]. Because bank loans/mortgages are the primary source of funding for the majority of the population, the availability of funding is an essential determinant of housing price levels [13]. In periods of economic prosperity marked by robust GDP growth, low inflation, and favourable interest rates, there is typically an upsurge in the demand for real estate properties [9]. According to [65], the allure of a buoyant economy often prompts individuals and investors to venture into the real estate market. Research done by [15] state that demographic factors such as population, ageing, and migration influence house prices. According to [67], a larger population correlates with higher real estate prices. Furthermore, if the proportion of the elderly population to the working population grows, property prices may be put under pressure. When a city's population grows, there is a greater demand for residential homes, which can drive up property prices [26]. A decrease in population, on the other hand, may result in decreased demand and potential price stagnation or reduction [26]. The unemployment rate affects the real estate market directly [15]. When unemployment is high, households struggle to make mortgage payments and frequently default on their loans. As a result, more homes are foreclosed on and finally sold at a loss by banks [18]. The fall in demand for housing causes price decreases, which leads to more defaults and price reductions [18]. Unemployment declines favour disposable income and cause agents to migrate to more affordable but also pricier homes [15]. The age distribution of the population has a direct impact on the types of properties in demand, for example, cities with a growing population of young professionals may see an increase in demand for modern apartments and smaller, urban-centric housing options [76]. Household composition changes can also influence housing preferences and, as a result, property prices [27]. For example, a growing proportion of single-person households may raise demand for smaller, one-

bedroom homes [69], while multi-generational households may choose larger, multi-bedroom houses [7]. Property costs are heavily influenced by the proximity of essential amenities such as schools, hospitals, retail centres, and public transit [59]. Properties near these amenities often attract higher prices as they provide residents with convenience and an increased quality of life [59]. For example, the proximity of well-regarded schools can be a significant driver of demand for properties in specific neighbourhoods [32]. Neighbourhood safety is an important concern for homebuyers, and low crime rates add to a location's overall desirability [63]. Property values are higher in areas with low crime rates and a general sense of security [71]. Homebuyers are willing to pay a premium for homes in safe and secure neighbourhoods [51]. The quality and availability of infrastructure within a neighbourhood, such as well-maintained roads, public transportation, and utilities, can improve an area's desirability and significantly impact property values [70]. According to [12], investments in infrastructure development and accessibility improvements are often associated with higher property prices in specific locations. The size and type of a property, whether a single-family home, an apartment, a condominium, or other housing, are key factors influencing its price [3]. Larger properties, such as large-family homes, typically command higher values than smaller units, such as apartments [74]. Properties that have been well-maintained and refurbished often have higher market values than those that require significant repairs or modifications [60]. First-time home purchasers often want move-in ready houses, which may justify higher prices [56]. Historical pricing trends may influence current property prices within a specific neighbourhood or location [17]. The demand for housing is intricately linked to the demographic structure of a population. As individuals age, their housing needs and preferences evolve, significantly influencing the housing market. Younger populations often prioritise different housing types compared to older groups, which can shape the overall demand for housing.[25]. In particular, research by [21] suggests that economies with a more significant proportion of older individuals tend to experience lower house prices. This trend is likely driven by the fact that older populations may require different types of housing, such as downsized or more accessible accommodations, reducing the demand for traditional family homes and, consequently, lowering prices in certain housing segments [25]. In addition to age, population growth is crucial in housing demand and price dynamics. Studies done by [25] indicate that a one per cent increase in population growth correlates with a 1.4 per cent rise in house price growth. This relationship highlights that as more individuals move into a given area, the demand for housing intensifies, exerting upward pressure on prices [25]. Such trends underscore the importance of both demographic changes and population shifts in shaping the housing market.

Zoning laws and land use policies establish the framework for property development and land use within a city and can influence property prices by limiting the types of structures developed in specific areas [39]; [34]. Zoning can limit or stimulate residential, commercial, or mixed-use development, influencing supply and demand for various property types [33]. Restrictive zoning restrictions, such as single-family zoning, reduce the supply of suitable land for new housing, raising the cost of new housing projects [29]. High property taxes can increase the overall cost of homeownership, thereby affecting affordability and property prices, while lower property tax areas may be more appealing to buyers as they minimise the ongoing financial burden of owning properties [14]. Local governments' rent control policies and affordability programmes can impact the rental market, which in turn influences property values [66]. Rent control policies may limit property owners' ability to generate rental revenue,

influencing property investment decisions and overall property values [6]. Local governments may offer tax credits or subsidies for affordable housing projects as property development incentives [48]. These incentives can impact housing availability and affordability, affecting property prices [37].

This section emphasises property size as a key determinant of real estate prices, with larger properties, such as single-family homes, generally commanding higher values than smaller units like apartments. The type of property, whether a house, apartment, or condominium, also significantly influences its market value. Additionally, demographics play a crucial role, as an increase in population typically drives up housing demand and prices, the same as the age of the population in the neighbourhood. While economic factors like GDP growth, interest rates, and inflation impact housing demand and pricing, this study primarily focuses on property size, demographics, condition of the property, and proximity to amenities, given the absence of historical real estate data in Vilnius municipality.

2.2 Traditional Real Estate Price Prediction Methods

Historically, real estate price prediction has relied on traditional statistical models, such as linear regression, hedonic pricing models, and autoregressive integrated moving average (ARIMA) models, which have been widely used. This section will overview these models and their application in real estate price prediction.

2.2.1 Hedonic Pricing Models

The hedonic pricing model has been extensively applied in the housing industry to analyze the impact of various factors on property prices. Numerous studies, [11], [43], [54] have utilized this model within the housing sector. These studies have been pivotal in estimating the value of non-observable attributes, such as proximity to neighbourhood amenities like hospitals and schools, environmental factors like air quality, and nuisances like airport noise levels.

By employing the hedonic pricing model, researchers have deduced household preferences for different housing characteristics and created housing price indices. The model typically establishes a relationship between house prices (the dependent variable) and various housing features (independent or explanatory variables), such as the number of bedrooms, location, or other attributes [54]. Mathematically, if we denote the house price as the dependent variable y and the housing characteristic as the independent variable x, the relationship is expressed through a regression equation in Equation 1 below:

$$y = \beta_0 + \beta_1 x + \epsilon, \tag{1}$$

Where:

- y is the house price,
- β_0 is the intercept,
- β_1 represents the coefficient of the housing characteristic x,
- ϵ is the error term accounting for other unexplained variations in house prices.

This regression model quantifies how different characteristics affect house prices, providing valuable insights for buyers and policymakers.

2.2.2 Linear Regression Models

Linear regression is another popular method in real estate price forecasting. It estimates the relationship between property prices and independent variables, such as the property's location, size, and age. While simple and interpretable, linear regression models often underperform when faced with nonlinearity and complex interactions in large datasets [30]. Additionally, linear models assume constant relationships between variables over time, which is a limitation in dynamic markets like real estate[5]. Linear regression is one of the simplest and most commonly used methods in real estate price prediction. It works by estimating the linear relationship between property prices (the dependent variable) and various independent variables, such as location, property size, number of rooms, and age of the building. This relationship is mathematically represented by the following equation:

$$P \;=\; \alpha \;+\; \beta_1 X_1 \;+\; \beta_2 X_2 \;+\; \ldots \;+\; \beta_n X_n \;+\; \epsilon, \tag{2}$$

where:

- P is the predicted property price (the dependent variable),
- $X_1, X_2, \ldots, X_n$ are the independent variables (e.g., Area, building age, neighbourhood),
- $\beta_1, \beta_2, \ldots, \beta_n$ are the regression coefficients capturing how each X_n contributes to the price,
- α is the intercept term,
- ϵ is the error term, accounting for deviations between predicted and actual prices.

As illustrated in Equation 2, the model assumes a linear combination of the independent variables to predict the property price. However, this assumption may not hold true in scenarios where the relationship between variables is inherently nonlinear.

The main advantages of linear regression are its simplicity and interpretability. It is easy to understand how each factor (e.g., number of bedrooms and proximity to schools) affects the overall property price. However, linear regression models often underperform when faced with large datasets containing complex, nonlinear interactions between variables [5].

In Ridge, Lasso, and Elastic Net Regression, the objective is to find the optimal set of coefficients $\boldsymbol{\beta}$ that minimizes the respective loss functions. The general form of the regularized objective function combines the **Residual Sum of Squares (RSS)** with a regularization term:

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta} \right)^2 + \lambda \cdot \text{Penalty}(\boldsymbol{\beta}) \right\}, \tag{3}$$

where:

- β_0 is the **intercept term**, representing the expected value of the dependent variable y when all independent variables $x_1, x_2, \ldots, x_p$ are zero,

- $\beta_1, \beta_2, \ldots, \beta_p$ are the **coefficients** corresponding to each independent variable $x_1, x_2, \ldots, x_p$, indicating the change in y for a one-unit change in the respective x_j, holding all other variables constant,
- λ is the **regularization parameter** that controls the strength of the penalty.
- Penalty($\boldsymbol{\beta}$) varies depending on the regression technique used.

In these methods, the goal is to find the values of $\boldsymbol{\beta}$ that minimize a *penalized* objective function, preventing overfitting by discouraging large or unnecessary coefficients. The objective function defined in Equation 3 combines the **Residual Sum of Squares (RSS)** with a regularization term that penalizes the magnitude of the coefficients.

In these methods, the goal is to find the values of $\boldsymbol{\beta}$ that minimize a *penalized* objective function, preventing overfitting by discouraging large or unnecessary coefficients.

Ridge Regression or L2 regularization addresses the issue of multicollinearity in simple linear regression by adding a penalty to the loss function, which helps produce more reliable coefficient estimates. Unlike simple linear regression, it allows for the retention of correlated predictors, thus maximizing the information extracted from the data. Ridge Regression seeks to minimize the following objective function:

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{4}$$

where:

- $\boldsymbol{\beta}$: The vector of coefficients to be estimated, including the intercept β_0,
- λ: The regularization parameter that controls the strength of the L2 penalty.

The objective function as expressed in Equation 4 comprises the **Residual Sum of Squares (RSS)** and the **L2 penalty** on the coefficients (excluding β_0 if not regularized)

While ridge regression is advantageous for handling multicollinearity, its inherent bias and the complexity of parameter selection can limit its applicability in certain contexts[41]. Researchers must weigh these factors when making a decision to use ridge regression in their analyses[41].

Lasso Regression Alternatively, L1 regularization also adds a penalty to the size of the coefficients but with a focus on sparsity. This technique can shrink some coefficients to zero, effectively performing feature selection.

Like linear regression, Lasso regression starts by calculating the sum of squared residuals. However, Lasso regression adds a penalty term to this calculation to discourage the coefficients of the independent variables from getting too large. This penalty term is the absolute value of the magnitude of the coefficients, hence the 'Least Absolute Shrinkage' in Lasso. The magnitude of this penalty term is governed by a parameter, typically denoted as λ (lambda).

Lasso Regression aims to minimize the following objective function in Equation 5:

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{5}$$

where:

- $\boldsymbol{\beta}$: The vector of coefficients, including the intercept β_0,
- λ: The regularization parameter controlling the strength of the L1 penalty.

The objective function includes the **RSS** and the **L1 penalty** on the coefficients, promoting sparsity by potentially setting some β_j to zero.

Elastic Net Regression combines both Ridge and Lasso regularization, incorporating both L1 and L2 penalties,minimizing the objective function in Equation 6. This method is beneficial when there are many highly correlated predictors.

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta} \right)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right) \right\}, \tag{6}$$

- $\boldsymbol{\beta}$: The vector of coefficients, including the intercept β_0,
- λ: The overall regularization parameter,
- α: The mixing parameter that balances the contribution of L1 and L2 penalties.

The objective function includes both the **RSS** and a combination of **L1** and **L2 penalties**, allowing for both variable selection and coefficient shrinkage.
Elastic Net regression is particularly effective when there are groups of correlated features. It can select groups of correlated features together, unlike Lasso, which may select one feature from a group and ignore the rest.
Linear, Ridge, Lasso, and Elastic Net Regression are widely used in house price prediction but face several limitations compared to machine learning methods. Simple linear regression assumes a linear relationship between predictors and the target variable, which often does not hold in the real estate market, where interactions and nonlinear patterns are common[42]. Ridge regression and Elastic Net handle multicollinearity effectively, while Lasso can struggle in cases of severe multicollinearity, especially with smaller datasets [20]. Regularized methods like Ridge and Lasso improve prediction accuracy over simple linear regression, particularly in high-dimensional settings. However, they often fall short of the predictive power of machine learning, which more effectively captures complex nonlinear relationships and variable interactions[44],[19]. Additionally, these methods may struggle with multicollinearity, especially when using dummy variables for categorical features[40]. While Ridge and Elastic Net can manage complex models, they may not fully exploit the potential of large datasets as adaptive machine learning algorithms can. However, machine learning approaches often require larger datasets and greater computational resources, which can limit their applicability in certain scenarios [19].

2.2.3 ARIMA Models

ARIMA models are commonly applied to time series data to capture trends and patterns in real estate prices over time. While ARIMA models can handle autocorrelation and seasonal effects, they struggle to incorporate external factors such as economic indicators and policy changes [61]. Moreover, ARIMA models are limited by their reliance on stationarity, making them less effective in capturing the complexities of real estate markets with rapidly changing conditions. Autoregressive Integrated Moving Average (ARIMA) models are commonly used for predicting time series data, making them a valuable tool for forecasting real estate prices over time. ARIMA models use historical price data to identify trends and patterns, capturing both short-term and long-term price movements. The ARIMA model is expressed as ARIMA(p, d, q), where:

- p is the number of lag observations (autoregressive terms),
- d is the degree of differencing (to make the series stationary),
- q is the size of the moving average window.

The general formula for ARIMA is given in Equation 7:

$$P_t = \alpha + \sum_{i=1}^{p} \phi_i P_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t, \tag{7}$$

where:

- P_t is the current value of the dependent variable (e.g., property price at time t),
- α is the intercept,
- ϕ_i are the coefficients for the autoregressive (AR) terms,
- θ_j are the coefficients for the moving average (MA) terms,
- ϵ_t is the error term at time t,
- p is the order of the autoregressive model,
- q is the order of the moving average model.

ARIMA models are effective in capturing autocorrelation and seasonality in real estate prices. For example, housing prices might follow a seasonal pattern due to weather or school enrollment cycles, and ARIMA can model such patterns accurately. However, one major limitation of ARIMA is that it assumes stationarity, meaning that the statistical properties of the time series (like the mean and variance) remain constant over time. This assumption may not hold in rapidly changing markets. Additionally, ARIMA models struggle to incorporate external factors like economic policies, interest rate changes, or sudden market shocks, making them less effective in dynamic environments [61].

2.3 Machine Learning Techniques for Real Estate Price Prediction

With the emergence of machine learning, more sophisticated models have been introduced for real estate price prediction, offering the ability to handle large datasets, nonlinear relationships, and complex interactions between variables. These models offer the ability to process large datasets and capture nonlinear relationships between variables, making them more effective than traditional methods [55].

2.3.1 Decision Tree

Decision trees are popular in real estate price prediction due to their simplicity and intuitive structure. A decision tree model works by splitting the dataset into smaller subsets based on the most significant features at each node. For example, a decision tree might first split the data based on location, then on property size, and finally on the number of bedrooms. The model then assigns a predicted value to each branch of the tree, eventually leading to a final prediction for the property price. The structure of a decision tree looks like this:

- Root Node: Represents the first feature that provides the most information gain (e.g., location).
- Internal Nodes: Represent the splitting rules based on other features (e.g., property size, age).
- Leaf Nodes: Represent the final prediction of the property price.

One of the advantages of decision trees is their ability to handle both numerical and categorical data, making them versatile for real estate datasets. However, decision trees are prone to overfitting, making them overly complex and too specific to the training data. This can lead to poor generalizability when applied to new data [64]. Pruning techniques and ensemble methods are often used to mitigate this issue.

2.3.2 Random Forests

Random Forest is an ensemble learning technique that improves decision trees by aggregating the predictions of multiple trees to reduce overfitting. Random Forests have been widely used in real estate price prediction due to their ability to handle large datasets, robustness, and capacity to capture nonlinear relationships [58]. Nevertheless, they can be computationally expensive and sometimes lack interpretability compared to simpler models. Random Forests are an ensemble learning method that extends decision trees by building multiple trees (a "forest") and averaging their predictions. Each tree in the Random Forest is trained on a random subset of the data using a technique called bootstrap aggregation, or "bagging," which helps reduce variance and overfitting. The formula for a Random Forest Regressor is as follows:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} f_i(x), \tag{8}$$

where:

- $\hat{y}$ is the final predicted value,

- N is the total number of trees in the forest,
- $f_i(x)$ is the prediction made by the i-th tree.

Each tree $f_i(x)$ as in Equation 8 is trained on a random subset of the data, and the final prediction is the average of the predictions of all trees in the forest.

Random Forest models are known for their robustness and ability to handle large datasets with complex interactions between variables. They are particularly effective in capturing nonlinear relationships, making them ideal for real estate price prediction. However, Random Forests can be computationally expensive due to the large number of trees involved, and they often lack interpretability compared to simpler models like linear regression [58].

2.3.3 Gradient Boosting Machines (GBM)

GBM is another popular machine learning model that builds successive models to correct errors made by previous models. Unlike Random Forests, which build trees independently, GBM builds trees sequentially. Each new tree is designed to correct the errors made by the previous tree, leading to progressively better predictions. It iteratively refines its predictions, making it highly effective in scenarios with complex data interactions, as often seen in real estate markets. Studies have shown GBM to outperform many traditional methods in real estate price forecasting due to its ability to handle non-linearity [45]. However, GBM models are sensitive to hyperparameter tuning and can be computationally intensive.

The formula for a Gradient Boosting Regressor is as follows:

$$\hat{y}^{(m)} = \hat{y}^{(m-1)} + \eta \cdot f_m(x), \tag{9}$$

where:

- $\hat{y}^{(m)}$ is the predicted value after the m-th iteration,
- $\hat{y}^{(m-1)}$ is the predicted value after the $(m-1)$-th iteration,
- η is the learning rate, controlling the contribution of each new tree,
- $f_m(x)$ is the function (usually a decision tree) added at the m-th iteration, trained on the residuals from the previous model.

In Equation 9, the function $f_m(x)$ is trained to minimize the residuals (or gradients) of the loss function, typically the mean squared error (MSE) in the case of regression.

Light Gradient Boosting Machine(LightGBM) is a powerful, high-performance machine learning framework that is particularly effective for large datasets and complex models. LightGBM is based on GBM, a technique that builds an ensemble of decision trees to make predictions. Unlike GBM, which uses a level-wise approach to build trees, LightGBM uses a leaf-wise approach, which grows the tree by selecting the leaf that leads to the largest reduction in the loss function [62]. However, the leaf-wise strategy can lead to trees with a higher depth, which may increase the risk of overfitting. To

mitigate this, LightGBM applies a maximum depth limit during tree growth to control the depth of the trees and reduce overfitting. LightGBM is known for its efficiency, speed, and ability to handle large-scale data, making it suitable for real estate price prediction and other regression tasks [10].

Extreme Gradient Boosting (XGBoost) is an optimized version of GBM that enhances the basic gradient boosting algorithm with additional techniques for improving speed, performance, and accuracy. XGBoost builds on the gradient boosting model by incorporating regularization (to prevent overfitting, making it robust against noisy datasets [62] and using second-order derivatives (Hessian) to optimize the tree-building process. This makes XGBoost not only faster but also more accurate than traditional gradient boosting.

The goal in XGBoost is to minimize the objective function by iteratively adding trees. The objective function is composed of two terms: the loss function and the regularization term, as defined in Equation 10.

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \ell(y_i, \hat{y}_i) + \Omega(f), \tag{10}$$

where:

- $\ell(y_i, \hat{y}_i)$ is the loss function (e.g., squared error for regression),
- $\Omega(f)$ is the regularization term that penalizes model complexity (penalizing the number of leaves and their weights).

The objective function defined in Equation 10 combines the **Residual Sum of Squares (RSS)** with a regularization term, enabling XGBoost to balance model fit and complexity effectively.

The final prediction for XGBoost is an additive model, where predictions are updated in each iteration by adding the output of a new tree, as shown in Equation 11.

$$\hat{y}_i = \hat{y}^{(0)} + \sum_{t=1}^{T} f_t(x_i), \tag{11}$$

where:

- $\hat{y}_i$ is the final prediction for the i-th sample,
- $\hat{y}^{(0)}$ is the initial prediction (usually the mean of the target values),
- $f_t(x_i)$ is the output of the t-th tree, trained on the residuals of the previous model.

GBM models, including variants like XGBoost and LightGBM, have shown high real estate price prediction accuracy due to their ability to model complex, nonlinear relationships between variables. However, these models are sensitive to hyperparameter tuning, and their computational costs can be high, especially for large datasets. Additionally, they are more prone to overfitting if not properly regularized [62].

2.4 Ensemble Methods in Real Estate Price Prediction

Ensemble techniques combine the predictions of multiple models to improve accuracy and robustness. The section discusses three ensemble methods used in real estate price prediction: bagging, boosting, and stacking.

2.4.1 Bagging (Bootstrap Aggregating)

Bagging is an ensemble technique that builds multiple models (often decision trees) and combines their predictions to reduce variance and improve accuracy. Random Forest is a prime example of a bagging technique used in real estate price prediction where multiple decision trees are trained, and their predictions are averaged. By averaging predictions across multiple trees, bagging helps mitigate the risk of overfitting, leading to more stable and reliable forecasts [47].

2.4.2 Boosting

Boosting is another ensemble method that builds models sequentially, with each new model correcting the errors made by its predecessor. Algorithms such as Gradient Boosting Machines (GBM), XGBoost, and LightGBM have become popular in real estate price prediction due to their high predictive accuracy. Boosting models are particularly effective in handling complex datasets and nonlinear relationships, making them suitable for the multi-dimensional nature of real estate markets [45]. However, they require careful tuning of hyperparameters to avoid overfitting and high computational costs.

2.4.3 Stacking

Stacking is an advanced ensemble method that combines multiple models by training a meta-model on their outputs. In real estate price prediction, stacking allows diverse algorithms, such as decision trees, boosting algorithms, and linear models, to achieve better predictive performance [72]. This approach benefits from leveraging the strengths of different models while compensating for their weaknesses. However, stacking models are more complex to implement and interpret than bagging and boosting methods.

2.5 Related Work

Numerous studies have focused on predicting real estate prices using machine learning models, each providing valuable insights into the methods and approaches that yield accurate results. This section reviews key research efforts contributing to real estate price prediction.

The research done by [77] involved a comprehensive study on predicting real estate prices using various regression techniques. Their work highlights the real estate sector's critical role in the broader economy, emphasising its contribution to job creation and wealth generation. The study's aim was to develop a predictive tool for the real estate market to estimate property prices based on house characteristics and their geographical locations. Given that the real estate market analysis is a vital component of strategic planning and investment decisions, their project sought to identify the most

influential factors affecting property prices. The findings of their work are intended to benefit both real estate consultants and prospective property investors by reducing investment risks. Using a dataset comprising 81 variables that described the sale of 1,460 residential properties in Ames, Iowa, USA, the authors applied four different regression models to predict housing prices. They also curated a modified dataset containing 10 selected variables from the original dataset for analysis. This study analysed the four regression models: Lasso Regression, Ridge Regression, Random Forest, and Extreme Gradient Boosting (XGBoost). The study evaluated the performance of these models using metrics such as R^2 score, root mean squared error (RMSE), and mean absolute error (MAE). Among the models tested, Random Forest outperformed the others, achieving an R^2 score of 0.8500, an MAE of 0.1132, and an RMSE of 0.1523. Moreover, the authors identified the most significant features influencing house prices across all four models, including OverallQual (a rating of the overall material and finish of the house), GrLivArea (the above-ground living area in square feet), and GarageArea (the size of the garage in square feet). These features emerged as the primary determinants for predicting real estate prices. This study offers valuable insights into the factors that affect house prices and highlights the effectiveness of Random Forest in predicting property values in the real estate market.

The study conducted by [46] focused on improving the accuracy of real estate price predictions using a combination of Deep Learning (DL) and Machine Learning (ML) techniques. Recognising the critical importance of accurate price predictions for stakeholders such as investors and developers, their research aimed to address the shortcomings of previous approaches in the field. The authors utilised the "House Prices 2023 Dataset" from Kaggle, which contains 168,000 entries of property data from Pakistan, making it one of the largest datasets in similar studies. Their methodology involved extensive data preparation, including feature engineering, and the implementation of various predictive models, such as Linear Regression, Gradient Boosting, Random Forest, Convolutional Neural Networks (CNN), and K-nearest neighbours (KNN). The models were evaluated using performance metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared, and accuracy. Among the models tested, KNN emerged as the best performer, achieving a lower RMSE of 13.79 and a higher R-squared value of 0.85, indicating improved predictive accuracy. Random Forest also produced notable results, achieving an accuracy rate of 80%. Despite these achievements, the study faced challenges, particularly in handling complex feature interactions, ensuring model scalability, and managing hardware resources efficiently. The research highlights areas for future improvement, including the need for enhanced computational efficiency and feature interaction handling. Overall, the study demonstrates the effectiveness of machine learning techniques, particularly KNN and Random Forest, in accurately predicting real estate prices, providing a solid foundation for further advancements in real estate market forecasting.

Another study by [8] compared the predictive performance of two models—Ordinary Least Squares (OLS) linear regression and Random Forest (RF)—for predicting apartment prices in Ljubljana, Slovenia. The dataset included 7,407 apartment transaction records from 2008 to 2013. The evaluation metrics used were R-squared, mean absolute percentage error (MAPE), and coefficient of dispersion (COD). The RF model performed better, achieving an R-squared of 0.82, a MAPE of 12 %, and a COD

of 0.15, whereas the OLS model had an R-squared of 0.65 and a higher MAPE of 18%. However, both models tended to overestimate lower prices and underestimate higher ones, with RF showing greater sensitivity to price variations. This study reinforced RF's ability to handle complex real estate datasets with non-linear relationships.

A different study carried out by [31] applied three machine learning models to a data sample containing about 40,000 housing sales transactions from Hong Kong over the past 18 years. The researchers evaluated the performance of Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Machine (GBM) models using RMSE, MAE, and R-squared. The RF model delivered an RMSE of 1.6, an MAE of 12.5, and an R-squared of 0.88, while GBM achieved an RMSE of 1.7, an MAE of 13.0, and an R-squared of 0.85. SVM performed the worst, with an RMSE of 2.2 and an R-squared of 0.72. This study underscored the benefits of ensemble methods, such as RF and GBM, for more accurate and robust predictions in real estate price forecasting.

The study by [73] used house prices and features in Ames, Iowa, from Kaggle with the aim of building a reliable house price prediction model. The data set contains 1121 records and 27 features after data preprocessing and exploratory data analysis, Linear Regression, XGBoost Regression, and Random Forest. This study found the overall quality of the house, its living area, and the total basement area to be the most influential factors affecting the housing prices in Ames, Iowa. Amongst the three models explored, Random Forest and XGBoost proved to be superior, achieving 0.8502 and 0.8803 R-squared values. Also, on exploring the Root Mean squared error, Random Forest has 35241 and XGBoost 30718

In a study by [1], advanced machine learning techniques were applied to predict house prices in Kuala Lumpur, Malaysia, using a dataset collected from Kaggle in 2019. The dataset contained 53,883 records and 12 features, including location, price, size, rooms, and proximity to amenities. Following data preprocessing and log transformation to normalise the price data, the study compared the performance of multiple linear regression (MLR), ridge regression (RR), LightGBM, and XGBoost models. Hyperparameter optimisation was conducted using GridSearchCV to enhance model efficiency. The evaluation metrics—Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Adjusted R-squared —indicated that XGBoost outperformed the other models, achieving the lowest MAE and RMSE values, and the highest Adjusted R-squared , which signified a better fit. The XGBoost model, with an R-squared 0.921 for training and 0.912 for testing were selected for deployment due to their superior predictive accuracy and consistency. Despite its strong performance, the model showed some challenges in pricing high-end locations, such as Mont Kiara and KLCC, where predictions sometimes deviated from actual values. The study concluded that XGBoost is effective in predicting housing prices and could be useful for future house buyers, investors, and policymakers in the real estate industry. Future work may explore additional features and extend the model to other regions in Malaysia to further enhance its predictive capabilities.

The research by [68] explored the performance of several machine learning models, including Random Forest, XGBoost, and LightGBM, as well as two hybrid methods—Hybrid Regression and Stacked Generalization Regression. the study utilized the Housing Price in Beijing dataset ehich contains more

than 300,000 records with 26 variables representing housing prices traded between 2009 and 2018. The evaluation was based on R-squared, MAE, and RMSE. Random Forest achieved an R-squared of 0.80, MAE of 11.7, and RMSE of 1.6 but suffered from overfitting. XGBoost and LightGBM performed comparably well, with LightGBM yielding an R-squared of 0.83, an MAE of 10.8, and an RMSE of 1.5, making it more time-efficient. The Hybrid Regression method outperformed all others, with an R-squared of 0.87 and an RMSE of 1.3. Stacked Generalization Regression delivered the highest accuracy overall, with an R-squared of 0.90 and the lowest RMSE of 1.2, despite its complexity. This study emphasised the potential of hybrid and stacking models in dealing with complex real estate data.

Source	**Dataset**	**Algorithm**	**Result**
[77]	Ames, Iowa; 1,460 records	Random Forest	R^2=0.8500, MAE=0.1132, RMSE=0.1523
[46]	House Prices Pakinstani 2023 Dataset, Kaggle ; 168,000 records	KNN	RMSE=13.79, R^2=0.85
[8]	Ljubljana, Slovenia; 7,407 records	Random Forest	R^2=0.82, MAPE=12%, COD=0.15
[31]	Hong Kong; 40,000 records	Random Forest	RMSE=1.6, MAE=12.5, R^2=0.88
[73]	Ames, Iowa, Kaggle; 1,121 records	XGBoost	R^2=0.8803, RMSE=30718
[1]	Kuala Lumpur, Kaggle 2019; 53,883 records	XGBoost	R^2=0.921 (train), 0.912 (test), Lowest MAE and RMSE
[68]	Housing Price in Beijing; 300,000 records; 40,000 transactions	Stacked Generalization Regression(LightGBM, Random Forest)	R^2=0.90, RMSE=1.2

Table 1: Summary of Results

These studies, summarized in Table 1, highlight the growing trend of using machine learning techniques in real estate price prediction, with a particular focus on ensemble methods such as Random Forest, XGBoost and LightGBM. While traditional models like simple linear Regression are still used, machine learning algorithms, especially ensemble techniques, are increasingly favoured for their ability to handle large datasets, non-linear relationships, and complex interactions between variables. The insights from these studies serve as a foundation for the present research, which aims to evaluate the performance of various machine learning models for predicting housing prices using data collected about Vilnius municipality.

2.6 Comparative Analysis of Machine Learning Techniques in Real Estate Price Prediction

Several studies have conducted comparative analyses of machine learning techniques for real estate price prediction. For instance, [62] compared Random Forest, LightGBM, and XGBoost models for real estate forecasting and found that a combination of LightGBM and XGBoost outperformed traditional methods, especially in capturing non-linear relationships between variables. Similarly, studies have shown that ensemble methods, particularly stacking, tend to perform better when a diverse set of base models are employed. However, the complexity of implementation and computation may be higher [68]. In a study by [72], the performance of traditional linear models, decision trees, and ensemble methods was compared using real estate data from urban areas in China. The results indicated that while decision trees and Random Forest performed well with structured data, boosting models such as LightGBM and XGBoost achieved the highest accuracy, particularly in datasets with complex and non-linear relationships.

2.7 Conclusion

The literature on real estate price prediction demonstrates the evolution of methods from traditional statistical models to advanced machine learning algorithms. While traditional models like linear regression and ARIMA are still used, they often fall short when faced with non-linear and complex datasets, which are common in real estate markets. Machine learning techniques, particularly ensemble methods such as bagging, boosting, and stacking, have proven to be more effective in handling the complexities of real estate price prediction. Key determinants of property prices were explored, emphasizing factors such as property size, type, demographics, condition, and proximity to amenities. Larger properties, such as single-family homes, tend to command higher market values than smaller units like apartments. Additionally, property type and demographic trends, including population growth and neighborhood age distribution, significantly influence pricing. While broader economic variables like GDP growth, inflation, and interest rates also impact housing markets, this chapter narrowed its focus to variables most relevant to Vilnius municipality, given the lack of historical real estate data.

This chapter presented the potential of machine learning in addressing the limitations of traditional models. It sets the stage for the methodology described in the next chapter, detailing the implementation of machine learning models, the process of data collection, and the framework for evaluating model performance.

3 Methodology

This chapter outlines the research methodology used in the project, which focuses on predicting real estate prices in Vilnius using machine learning models. The methodology follows a systematic approach that involves data collection through web scraping, data preprocessing, feature engineering, model selection, model evaluation, and hyperparameter optimization. Each step is crucial for ensuring the predictive models are accurate, robust, and applicable to real-world scenarios.

3.1 Data Collection

The primary dataset used in this research was scraped from a leading real estate platform in Vilnius. Web scraping was chosen due to the lack of publicly available datasets on real estate transactions that provide up-to-date and detailed property characteristics.

3.1.1 Web Scraping Procedure

Web scraping was implemented using Python, employing the *Selenium* and *Pandas* library to programmatically extract data from the real estate website *Aruodas.lt* (`https://m.en.aruodas.lt/butu-nuoma/vilniuje/`). The process began with identifying a suitable primary website offering comprehensive real estate listing on apartment sales. Python script was developed using Selenium to navigate the website dynamically, interact with property listings, and handle pagination, ensuring that all relevant data was retrieved. Once the content was accessed, *Pandas* was utilized to parse extracted data into a CSV format, making it ready for further pre-processing and analysis.

The extracted dataset(Apartment data) consists of approximately 3,000 real estate records with various attributes related to properties in Vilnius. Each record includes details such as:

- Property location (city, neighbourhood)
- Property description (number of rooms, floor number, total floors, area)
- Year of construction or renovation
- Building type, heating system, and energy efficiency class
- Distances to nearby public amenities (kindergartens, schools, shops, transport stops)
- Price (the target variable for prediction)

This data set provided a comprehensive foundation for the machine learning tasks undertaken in this study.

3.1.2 Demographics Data

In order to enrich the real estate dataset, demographic information was integrated from the *Vilnius Open Data* repository (`https://github.com/vilnius/gyventojai`). The demographic data included the following attributes:

- **Demographic Attributes:** Birth year, birth country, gender, marital status, and number of children.
- **Geographic Identifiers:** Neighbourhood, street, Neighbourhood ID, region code, street code, and street ID.

The demographic data was aggregated at the street level and joined with the real estate dataset. This integration allowed for a richer analysis, providing a deeper understanding of how demographic trends correlate with real estate pricing in Vilnius.

3.1.3 Ethical Consideration

The apartment data was collected from a publicly available real estate website. The scraping was carried out in batches so as not to overwhelm the server with the scraper bot, and no data manipulation was carried out during scraping. No personal or sensitive information was used for the research, as the unique number of the object was deleted during the data cleaning process. Demographic data was sourced from Vilnius Open Data, freely available on GitHub. This data was aggregated, ensuring anonymity and eliminating any traceability to individuals or specific entities.

3.2 Data Cleaning and Preprocessing

Data preprocessing is critical in transforming raw data into a suitable format for machine learning models. The following operations were performed:

3.2.1 Unit Standardization

To maintain consistency and facilitate numerical computations, units across relevant features were standardized. Specifically:

- All distances in the proximity-to-amenities fields were converted to meters, harmonizing mixed entries of meters and kilometres.
- Units were removed from numeric fields such as *Price* and *Area*, enabling their conversion to integer values.

3.2.2 Handling Missing Data

After collecting the data, some records were found to contain missing values in key attributes, such as "Heating System," "Building Energy Efficiency Class," and distances to nearby public amenities. Different strategies were employed to address these gaps based on the type of data involved. Features with over 75 % missing values were excluded from the analysis due to their lack of informative values. For numerical features, particularly the distances to public amenities, missing values were imputed using the respective neighbourhood's median. This approach preserved locality-specific patterns, minimizing the risk of introducing bias.

3.2.3 Removing Districts in the Outskirts of Vilnius

The outskirts of Vilnius are characterized by a significantly lower number of real estate listings compared to central districts. This discrepancy arises from the predominant type of housing in these areas—detached houses rather than apartments. Since the primary focus of this study is apartment prices, including these districts could introduce biases due to the differing dynamics of housing types.

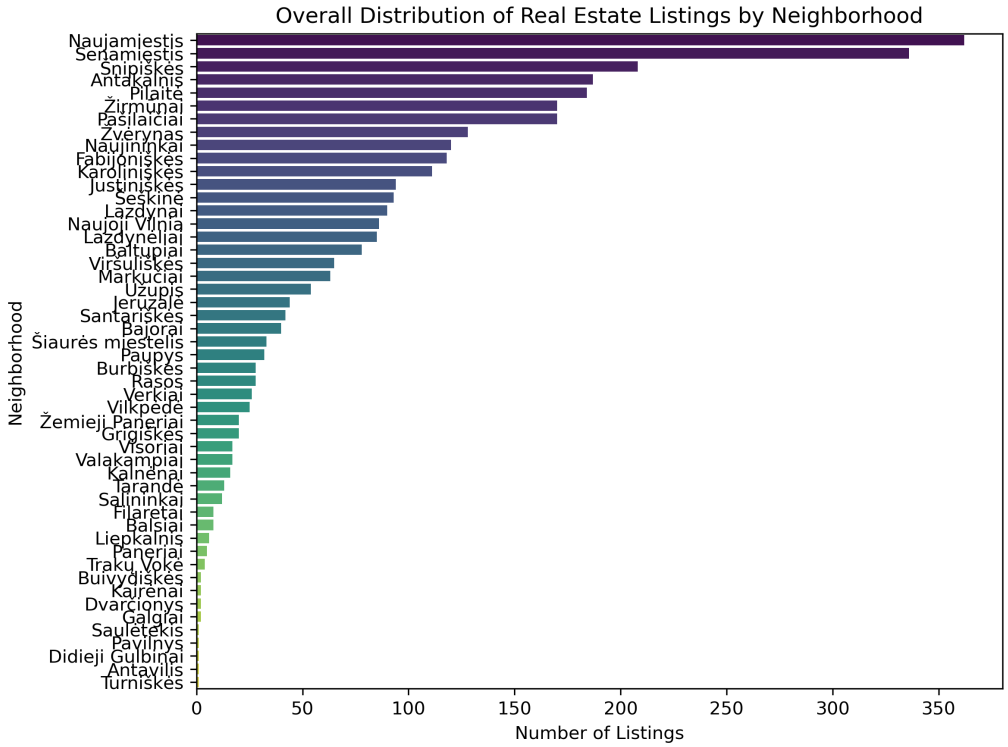


Figure 1: Overall distribution of real estate listings by neighborhood.

As shown in Figure 1, the central districts such as *Naujamiestis*, *Senamiestis*, and *Šnipiškės* exhibit the highest number of real estate listings, reflecting their urban nature and higher apartment density. In contrast, peripheral neighborhoods like *Didieji Gulbinai*, *Antavilis*, and *Turniškės* show significantly fewer listings due to their predominant housing type being detached houses rather than apartments.

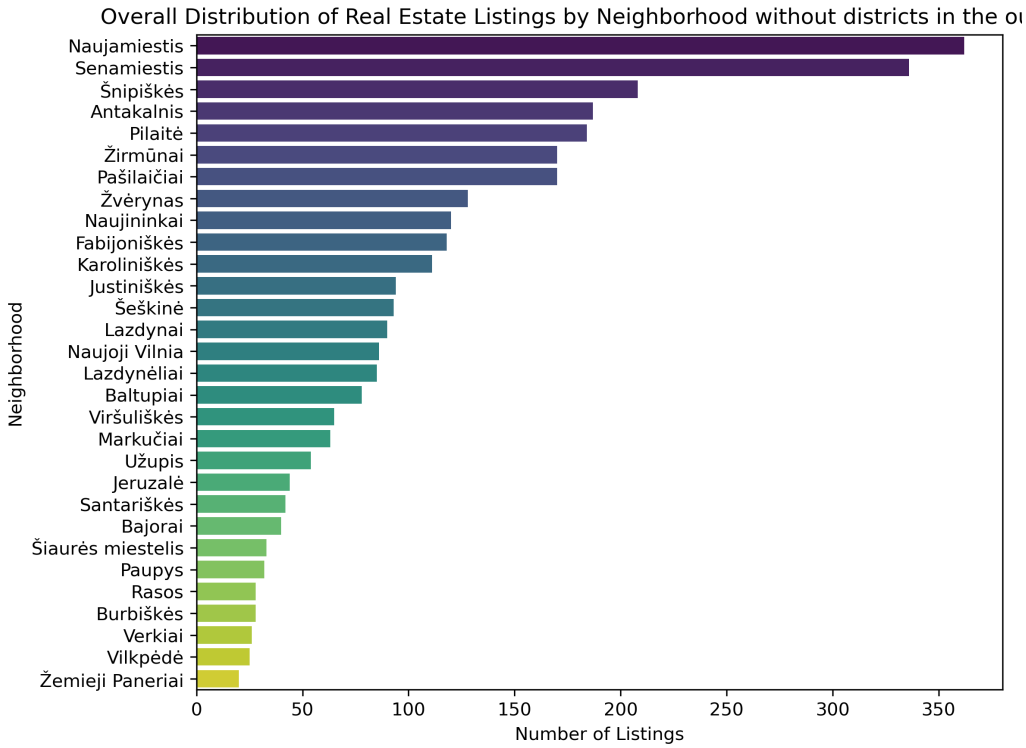


Figure 2: Distribution of real estate listings by neighborhood after removing districts in the outskirts.

To improve the robustness of the dataset and eliminate noise, neighbourhoods with very low listing counts were excluded. The updated chart in Figure 2 demonstrates the distribution of listings after removing these districts, which better concentrates the dataset on areas with sufficient data for meaningful machine learning analysis. This adjustment ensures that the machine learning models focus on neighborhoods with higher data availability, enabling more accurate predictions and reducing potential biases introduced by data sparsity in outskirt districts.

3.2.4 Dealing with Outliers

Outliers can significantly distort the distribution of data and adversely impact the performance of machine learning models. Therefore, identifying and removing outliers is a critical step in the data preprocessing pipeline. During the data cleaning phase, certain extreme values for features like property prices and area were identified as potential outliers. This section describes the approach used to detect and handle outliers in the *Price* feature.

Identification of Outliers: To better understand the distribution of the *Price* variable, a Q-Q plot was generated to compare the distribution of the *Price* variable with a theoretical normal distribution.

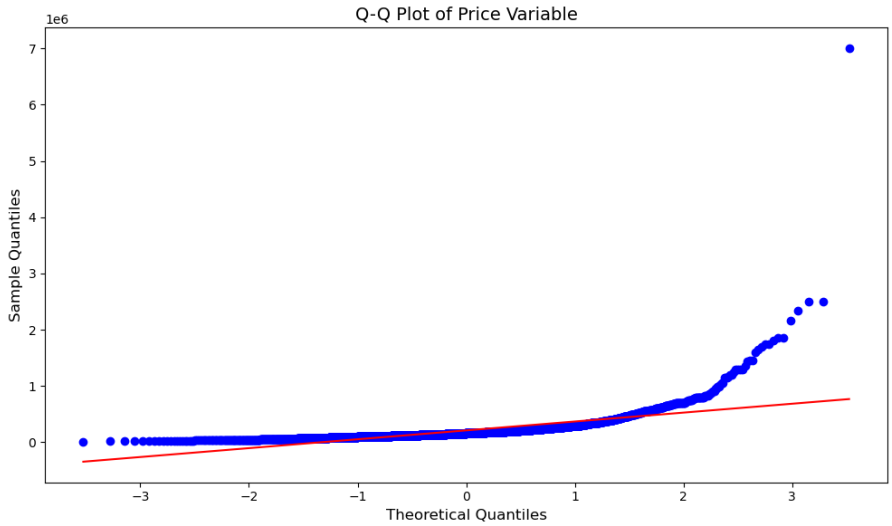


Figure 3: Q-Q Plot of Price Variable.

In the Q-Q plot, the X-axis represents the theoretical quantiles of a standard normal distribution, showing what the property prices would look like if they perfectly followed a normal distribution. The Y-axis displays the actual quantiles from prices feature, allowing for a direct comparison. The red diagonal line serves as a reference, representing the ideal scenario where the sample quantiles align perfectly with the theoretical normal distribution. Observing Figure 3 , clearly many points deviate from this red line, particularly in the upper tail indicating that the price feature does not follow a normal distribution , suggesting the presence of outliers.

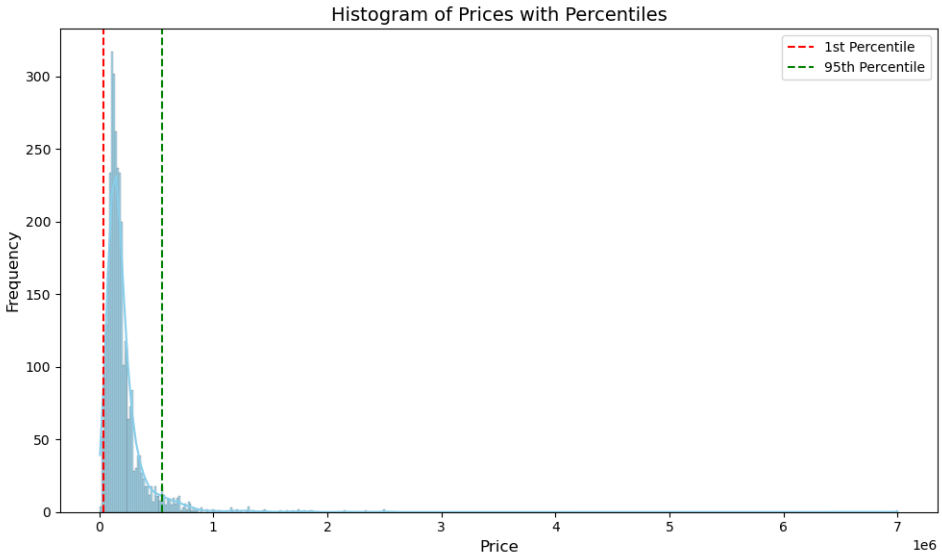


Figure 4: Histogram of property prices with percentile thresholds.

The analysis of the *Price* variable shown in Figure 4 revealed a highly right skewed distribution with extreme values on right end. To systematically identify outliers, the 1st percentile (1%) and 95th percentile (95%) were chosen as thresholds:

- **Lower Bound:** Prices below the 1st percentile.
- **Upper Bound:** Prices above the 95th percentile.

These bounds were computed using the `quantile()` function in python, which is robust to extreme values and provides a reliable estimate of percentiles in skewed distributions.

Removal of Outliers: To clean the dataset, observations with price values below the 1st percentile or above the 95th percentile were removed. This process preserved the central 94% of the data, while excluding extreme cases that could skew the analysis. This cleaned dataset formed the basis for further analysis.

3.3 Encoding Categorical Variables

Machine learning algorithms operate predominantly on numerical data, making it imperative to convert categorical variables into a numerical format before modelling. In this study, catergorical features such as *neighbourhood*, *street*, *Building Type*, and *Equipment* were transformed into numerical formats using encoding techniques based on the nature of the feature.

- **Label Encoding:** Label encoding assigns a unique integer to each category, ensuring a compact representation of the data while maintaining simplicity. Label encoding was applied to features such as *neighbourhood*, *building type*, and *street*.
- **Ordinal Encoding:** Used for the *equipment* feature, which has a clear order(e.g.Not equipped, Partially equipped ,Fully equipped).Ordinal encoding ensures that the resultant encoded feature retained the intrinsic ranking, there by allowing models interpret this feature effectively.

3.4 Feature Scaling

The dataset used for predicting real estate prices includes numerical characteristics with varying units and scales, such as the area of the property measured in square meters, the distances to the public facilities measured in metres, and the demographics statistics. These differences in scale can significantly impact the performance of machine learning models, as certain algorithms, especially those relying on distance-based calculations like Gradient Boosting or Linear Regression, can be sensitive to the magnitude of the features. Larger values, such as distances, can dominate smaller ones, like the number of rooms, potentially skewing the model's predictions. To address this issue and ensure that all numerical features are treated equally by the algorithms, StandardScaler from the sklearn.preprocessing module was applied. StandardScaler works by transforming the data so that each feature has a mean of 0 and a standard deviation of 1. This process, known as standardization, adjusts the distribution of each numerical feature, making them comparable in terms of scale without altering their underlying relationships. For example, after applying StandardScaler, the property area and distance to the nearest school, though measured in different units, will have similar numerical ranges, enabling the machine learning models to learn more effectively from the data. By applying StandardScaler, the model is better equipped to handle features of different magnitudes, improving convergence during training, and

potentially increasing predictive accuracy. Standardizing the data also helps prevent certain features from having an outsized influence on the model, ensuring a more balanced approach to prediction [6].

3.5 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process that involves investigating and summarizing the main characteristics of a dataset. The primary goal of EDA is to gain insights into the data's structure, distribution, and patterns before applying more complex modeling techniques. It typically includes visualizing the data through graphs such as histograms, box plots, scatter plots, and heatmaps, as well as calculating summary statistics like means, medians, and correlations between variables. By performing EDA, data scientists can detect outliers, handle missing data, identify relationships between variables, and ensure data quality, all of which help guide the choice of further modeling techniques and preprocessing steps. Ultimately, EDA provides an understanding of the dataset's underlying structure, which is essential for making informed decisions during the modeling process.

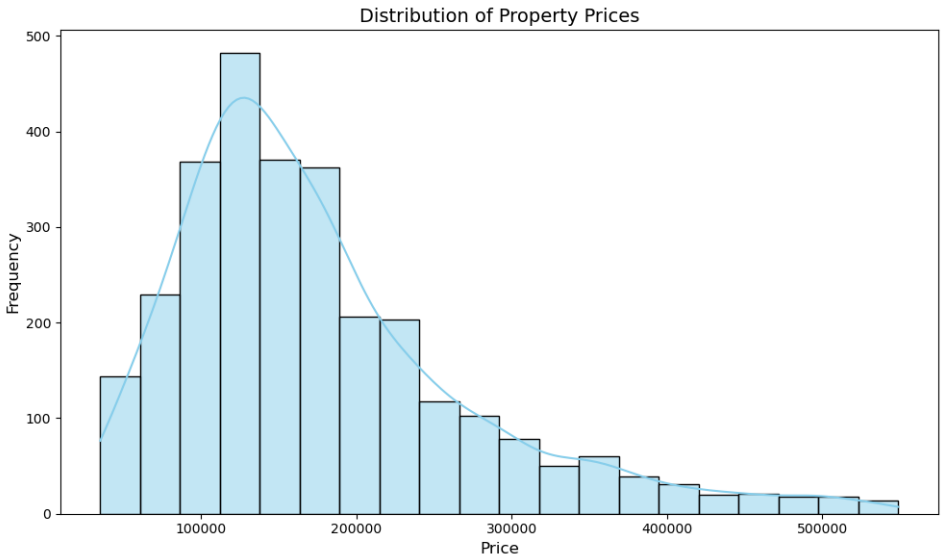


Figure 5: Distribution of property prices.

The distribution of property prices, as shown in the Figure 5 , highlights a right-skewed pattern. Property prices are distributed over a wide range, with notable peaks between 100,000 and 200,000. A gradual decrease in frequency is observed as prices rise, with a smaller but notable cluster around 200,000 and 300000, followed by a tapering off toward 300,000 and above. The frequency of properties priced above 300,000 is relatively lower, suggesting that higher-priced properties are less common in the data set. The smooth curve overlaid on the histogram suggests an overall trend in property prices, with the data showing moderate variation, but no extreme skewness in either direction. This distribution pattern could be indicative of a segmented market, with a large proportion of properties falling within midrange price categories and fewer properties at higher price points.

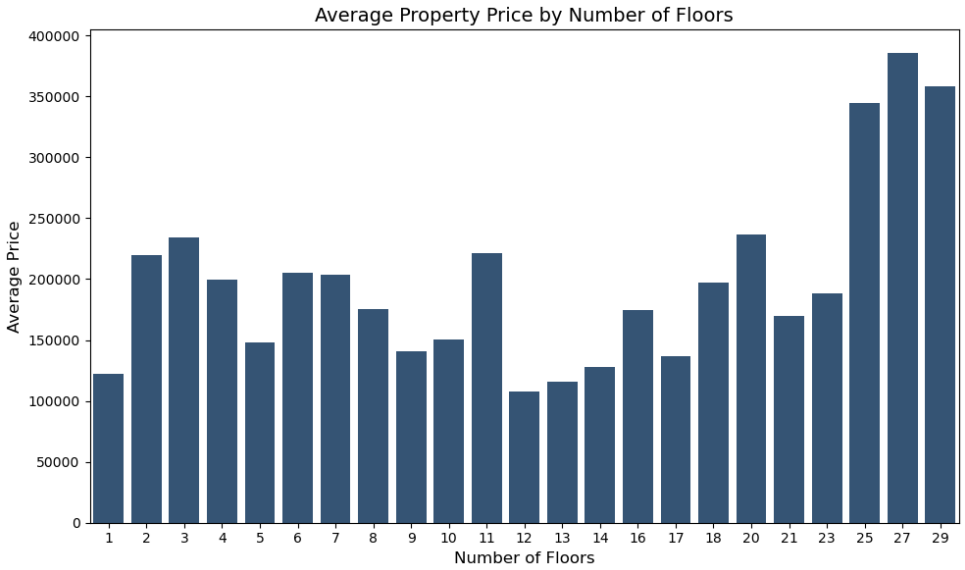


Figure 6: Average Price distribution based on number of floors

Figure 6 displays the relationship between property prices and the number of floors in the building. Properties with 27 floors are priced the highest, with an average of 385500, followed closely by those with 29 floors at 358,000 and 25 floors at 345,000. This is likely an indicator that high-rise buildings are more desirable or seen as premium properties.

Interestingly, in contrast to the highly priced high-rise building buildings with 2 and 3 floors also show relatively high average prices, at 219,798 and 234,329, respectively.

Mid-rise buildings, ranging between 5 and 18 floors, tend to fall in a more moderate price range, with averages generally between 148,199 and 196,880. For instance, buildings with 9 floors average 140,758, while those with 12 floors and 14 floors average 116,000 and 127,833, respectively. These prices reflect a more affordable segment of the market, though they still exhibit variability depending on the specific number of floors.

Notably, buildings with 21 floors begin to show a clear increase in average prices, reaching 169,500, and prices continue to rise sharply for properties in buildings with 25 or more floors. This trend suggests a complex relationship between the number of floors and property prices, where both low-rise and very high-rise buildings are more valuable compared to mid-rise buildings.

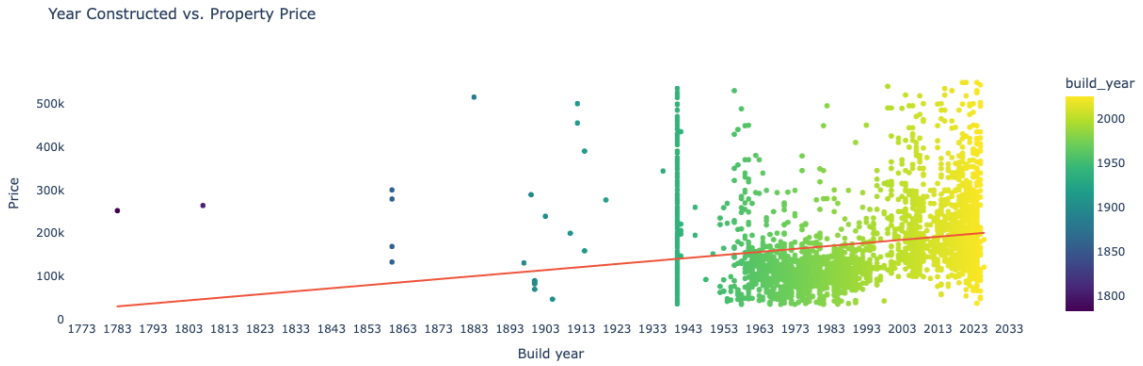


Figure 7: Construction year over Property price

The scatter plot in Figure 7 shows the relationship between the year of construction and the price of properties, with the color gradient representing the build year. Older properties (built before 1900) are displayed in darker colors, while more recent constructions (post-2000) are shown in brighter yellow hues. The red trend line indicates a positive correlation between the year of construction and property prices, suggesting that newer buildings tend to command higher sale prices. The dataset set has very few listings built before the 1953. Apartments built between 1953 and 1993 show a significant variability in prices with most of the listings priced under 200,000. Properties constructed post 1993 especially post 2003 exhibit a striking increase in value with many exceeding 300,0000. The trend suggests that newer properties are valued higher, likely due to factors such as modern amenities, better construction standards, and more desirable locations. The general increase in prices with newer buildings highlights the premium placed on recent developments in the real estate market.

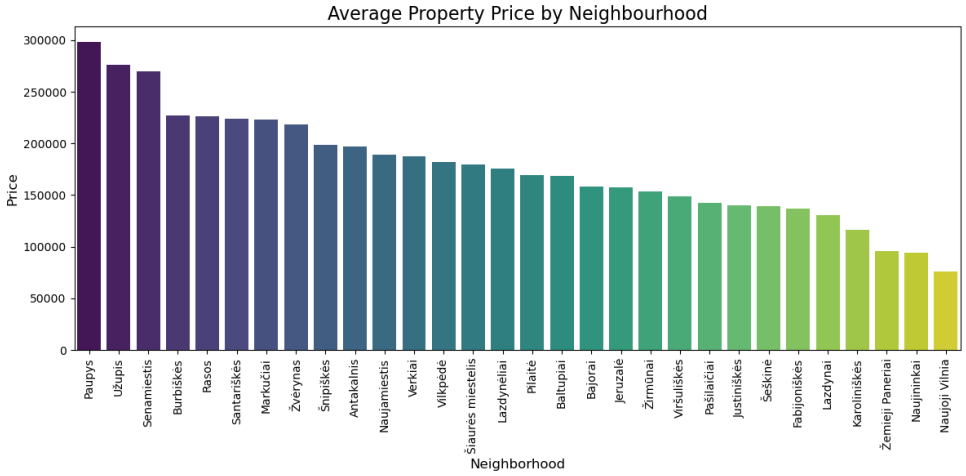


Figure 8: Average property prices distribution per Nerighbourhood

The bar chart in Figure 8 illustrates the average sale price of properties across different neighborhoods in Vilnius. Neighborhoods such as Paupys, Užupis, and Senamiestis show the highest average property prices, with values exceeding 250,000 , indicating that these areas are likely to be more affluent or desirable. In contrast, neighborhoods like Žemieji Paneriai, Naujininkai and Naujoji Vilnia display significantly lower average property prices, falling below the 100,000 mark, which suggests these

areas might be less in demand or offer more affordable housing options. Other neighborhoods, such as Antakalnis, Naujamiestis, Verkiai, Vilkpėdė, Šiaurės miestelis, and Lazdynėliai have moderate pricing, averaging between 175,000 and 200,000. The variation in property prices across neighborhoods highlights the diverse real estate market in Vilnius, where certain districts command premium prices, likely due to factors such as location and the construction year of properties in the neighbourhood.

Figure 9: Maximum and Mininum property Prices per Neighbourhood

Figure 9 showcases the minimum and maximum sale prices across different neighborhoods in Vilnius. It highlights significant price variability within each neighborhood. Several neighborhoods, including Markučiai, Antakalnis, Užupis, Senamiestis, Paupys, Naujamiestis, and Žvėrynas, have maximum property prices exceeding 500,000. However, with the exception of Paupys, these neighborhoods show a substantial gap between their minimum and maximum prices, with minimum prices ranging between 35,000 and 50,000. Paupys, a high-end neighborhood, exhibits a smaller price gap, with property prices starting at 149,999, reflecting its exclusive and upscale nature. Other neighborhoods, such as Burbiškės (minimum 115,000, maximum 410,000) and Santariškės (minimum 112,000, maximum 415,000), exhibit moderate price ranges, suggesting a balance between affordability and premium housing options.

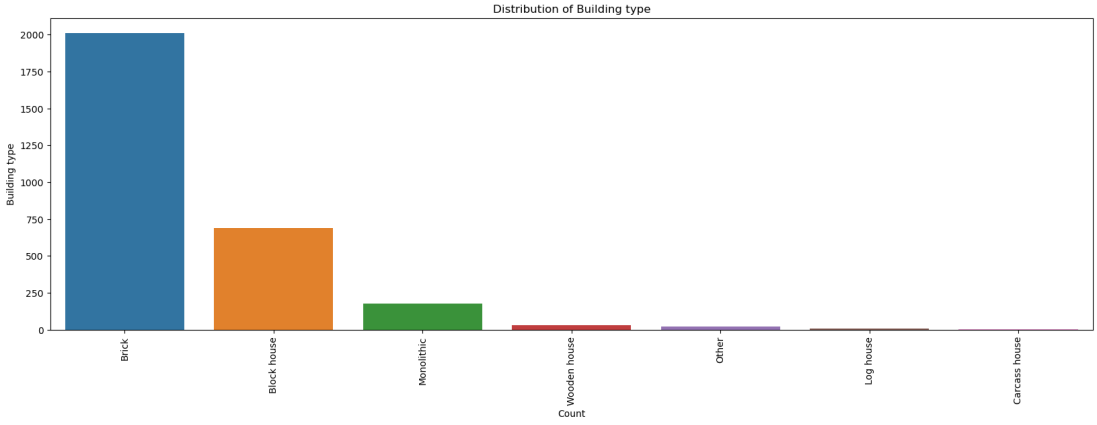


Figure 10: Distribution of Building type across apartment listing

Figure 10 illustrates the count of houses sold based on different building types in Vilnius. The data

shows that brick houses are by far the most popular type of building, with over 2,000 units in the listings, significantly outpacing other building types. Block houses come in second, with around 600 apartments , while monolithic buildings account for fewer listings , with less than 250 units. Lastly, the other categories which includes carcass house, log house , wooden house and others represent the smallest group accounting for only 0.02% of the listings This distribution highlights the dominance of brick houses in the market, indicating their popularity and potentially greater availability in Vilnius. The relatively lower counts for block and monolithic structures could reflect either their niche market status or their limited supply compared to brick houses.

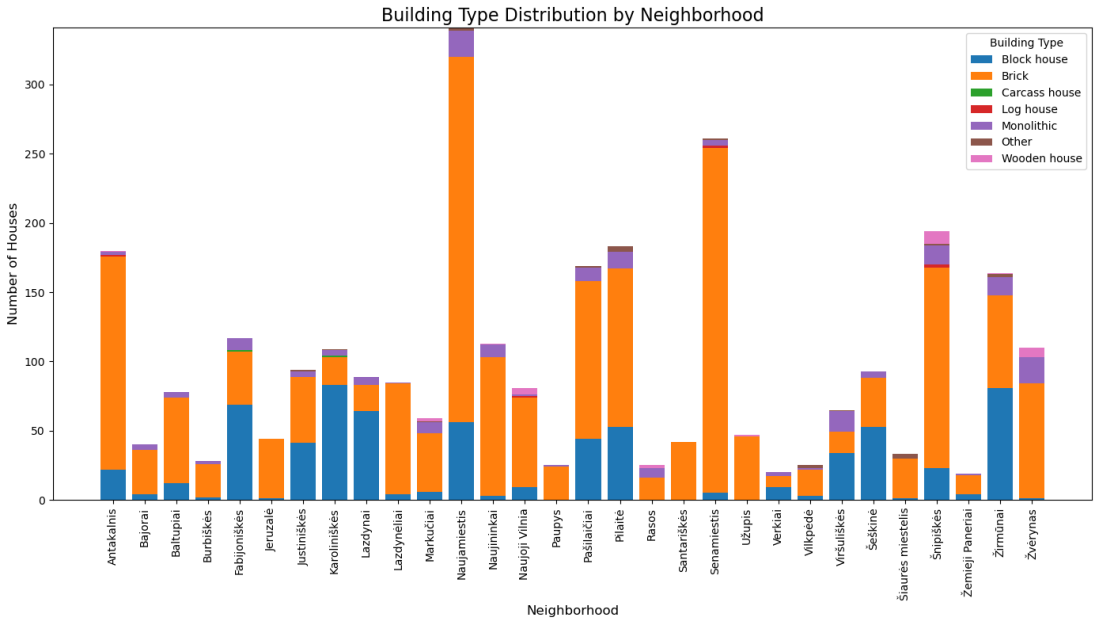


Figure 11: Building Type Distribution by Neighbourhood

Figure 11 illustrates the distribution of building types across various neighborhoods in Vilnius. Brick houses dominate the landscape in most neighborhoods, particularly in Santariškės, Užupis, Jeruzalė, Paupys, Lazdynėliai, and Senamiestis, where they account for more than 90% of property listings. This dominance suggests that brick houses are either the most readily available or the most preferred option in these areas.

Block houses, on the other hand, are notably prevalent in Karoliniškės and Lazdynai, where they make up the largest share of apartment listings. This indicates a preference for this building type in these neighborhoods, likely due to historical construction trends or specific housing demands.

Monolithic buildings, which are considered more modern construction types, appear most frequently in Žvėrynas and Naujamiestis, reflecting the contemporary architectural styles and possibly higher-end developments in these areas.

Together, Figures 11 and 19 highlight the distinct architectural and construction preferences across neighborhoods in Vilnius. For instance, while the prevalence of brick houses in many neighborhoods underscores their enduring popularity or widespread availability, areas like Žirmūnai and Justiniškės exhibit a broader diversity of building types. This diversity suggests that these neighborhoods may cater to a wider range of buyer preferences and housing needs.

Figure 12: Scatter plot showing the relationship between property area and Property price

Figure 12 shows how the area of the property relates to the property price. The scatter plot highlights a clear trend: as the size of a property increases, its price also tends to rise. This relationship is captured by the red regression line, which provides the best fit for the data and emphasises the positive correlation between area and price. The majority of apartments in the data used in this research fall within the range of 50 to 100 square metres, with prices typically between 100,000 and 250,000 Euros. However, there are a few listings exceeding 150 square metres. The shaded area around the regression line represents the 95% confidence interval, offering an estimate of the variability in the relationship. While the trend line indicates the general direction, the scatter plot shows some variability, especially for smaller properties under 100 square meters.

Overall, this plot confirms that property area is a key factor in determining price, as is often true in real estate markets. Larger properties tend to cost more, but the data also highlights the need to consider other variables to fully understand pricing differences, as some apartments within 50 to 100 square metres are high priced.

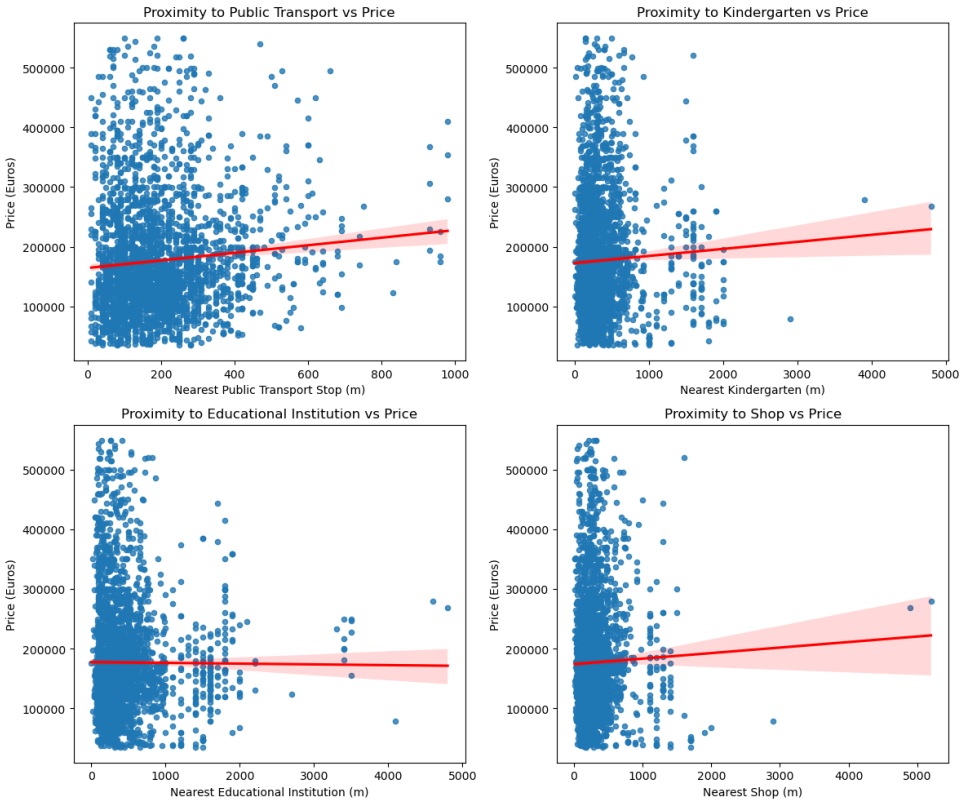


Figure 13: Scatter plots showing the relationship between property price (in Euros) and proximity (in meters) to key amenities

Figure 13 explores how the proximity of properties to key amenities, including public transport stops, kindergartens, educational institutions, and shops, relates to their prices. Each plot includes a red regression line to show the general trend and a shaded confidence interval to indicate the uncertainty around the trend. These plots indicate that there is no clear or strong correlation between the proximity to shop, kindergarten, public transport. In the dataset used for this research, most apartments have at least one public transport stop within a 400-meter radius. Similarly, the majority of apartments are located within 1 kilometer of a kindergarten, educational institution, and shop suggesting that accessibility to these amenities is relatively uniform across the dataset, which may explain the lack of significant variation in property prices based on proximity.

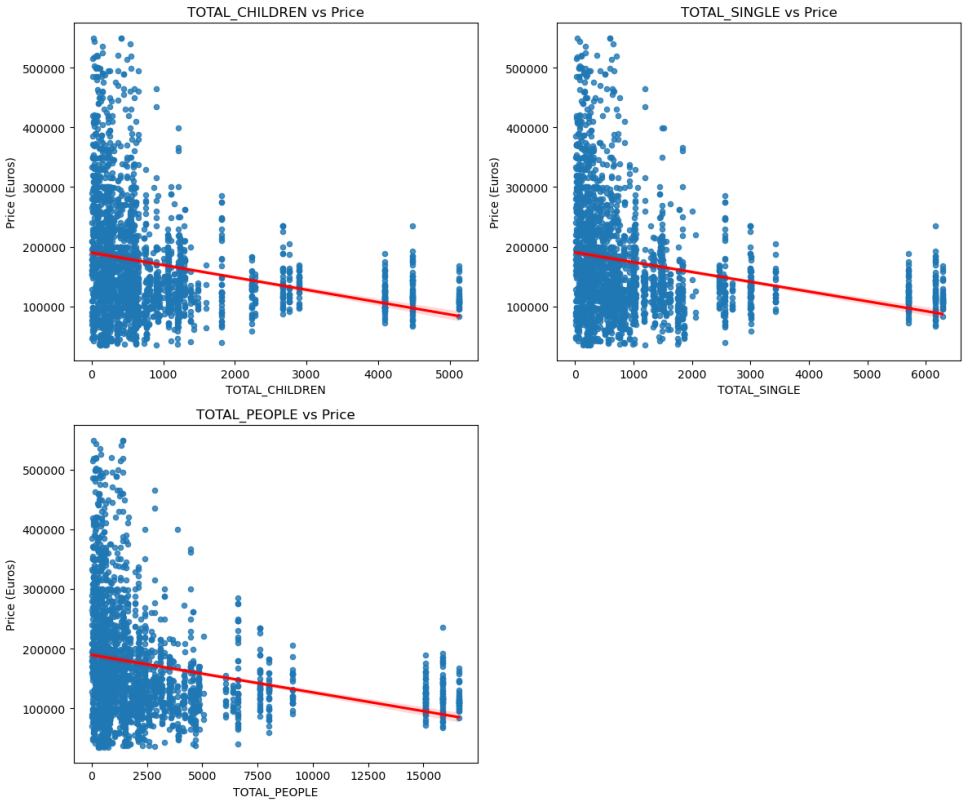


Figure 14: Scatter plots showing the relationship between demographic features (total number of children, single individuals, and total population) and property prices

Figure 14 illustrates the relationship between property prices and three demographic variables: total number of children, total number of single individuals, and total population in the neighbourhood where the apartment is situated. All three plots show a negative correlation between property prices and the demographic variables. Property prices tend to decrease as the number of children, single individuals, or total population increases. Overall, apartment prices are lower in less populated neighbourhoods, and higher population density might be associated with neighbourhoods where housing is more accessible but less premium, which can lower average property prices.

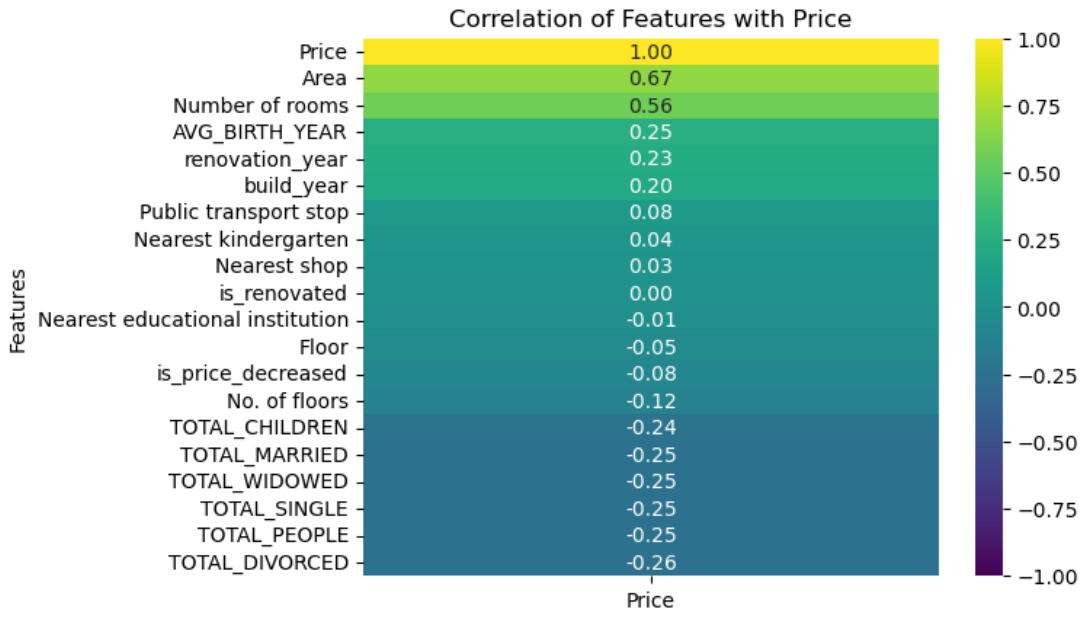


Figure 15: Correlation heatmap of numerical features

The heat map in Figure 15 shows how different features correlate with apartment prices, using a colour scale that ranges from -1.0 to $+1.0$. Notably, the size of an apartment—reflected by *area* (0.67) and *number of rooms* (0.56)—emerges as a key driver of price, suggesting that larger apartments with more rooms tend to be more expensive.

Features such as *renovation year* (0.23) and *build year* (0.20), which are related to the property's age and condition, also have positive correlations, indicating that newer or recently upgraded apartments may command higher prices. In contrast, features that indicate an apartment's proximity to amenities like *public transport* (0.08), *kindergartens* (0.04), and *shops* (0.03) show minor positive effects, suggesting these factors, while important, are not as influential as size. Interestingly, the feature *is_renovated* (0.00) has almost no effect, implying that when a renovation took place might matter more than whether it happened at all. Additionally, features like *floor* (-0.05) and *is_price_decreased* (-0.08) have minimal influence on the price.

Looking at the demographic features, such as *TOTAL_DIVORCED* (-0.26), *TOTAL_SINGLE* (-0.25), *TOTAL_PEOPLE* (-0.25), and *TOTAL_MARRIED* (-0.25), we see a clear negative association with price, indicating that areas with higher numbers of these populations tend to have lower apartment prices. Overall, these trends highlight how price is shaped more strongly by apartment size and upkeep factors, while demographic makeup exerts a measurable, though negative, pull on prices.

3.6 Feature Engineering

Feature engineering was a critical component of this study, involving deriving new features to enrich the dataset and enhance predictive performance. Key engineered features include measures of building age, time since renovation, floor ratio, among others. A detailed overview of the newly engineered features can be viewed in Table 11. The effects of these newly engineered features are further discussed in Section 4.9.

3.7 Model Selection

This research applied a range of machine learning algorithms to predict property prices, leveraging both linear models and ensemble techniques. Linear Regression was used as the baseline model for comparison. To enhance performance and address potential overfitting, regularized linear models—including Lasso, Ridge, and Elastic Net—were built. In addition to these linear models, various ensemble methods were implemented to capture complex patterns in the data. These included Random Forest, LightGBM, XGBoost which frequently outperformed other models in the studies reviewed in addition to Gradient Boosting, and Decision Trees. The models were compared by calculating the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), to determine the most effective approach for property price prediction.

3.8 Hyperparameter Tuning

To optimize the performance of the predictive models, hyperparameter tuning was systematically conducted using a *grid search* approach. Grid search is a comprehensive method that exhaustively explores a manually specified subset of hyperparameters to identify the optimal settings for each model. This process ensures that each model operates under its most effective configuration, thereby enhancing predictive accuracy and generalization capabilities.

Grid Search Methodology For each model, key hyperparameters that significantly influence performance and complexity were selected based on their impact on model bias, variance, and computational efficiency. For tree-based models, hyperparameters such as tree depth and the number of estimators were tuned to control model complexity and prevent overfitting. In boosting algorithms like Gradient Boosting, XGBoost, and LightGBM, learning rates and subsampling ratios were adjusted to balance bias and variance effectively. Additionally, regularization parameters were tuned for linear models to mitigate multicollinearity and enhance feature selection. The specific hyperparameters tuned for each model, along with the ranges of values considered, are summarized in Table 2.

Model	Hyperparameters
Decision Tree	max_depth: [None, 3, 5, 7, 9, 10, 13, 15, 18] min_samples_split: [2, 4, 5, 7, 9, 10] min_samples_leaf: [1, 2, 3, 4, 5, 7, 9] max_features: [None, 'sqrt', 'log2'] criterion: ['absolute_error']
Random Forest	n_estimators: [100, 200, 500, 1000] max_depth: [None, 3, 5, 7] min_samples_split: [2, 5, 10] min_samples_leaf: [1, 2, 5]
Gradient Boosting	learning_rate: [0.05, 0.1, 0.2, 0.5] n_estimators: [100, 200, 500] max_depth: [3, 6, 10, 15] min_samples_split: [2, 5, 10]
XGBoost	learning_rate: [0.05, 0.1, 0.2, 0.5] n_estimators: [100, 200, 500] max_depth: [3, 6, 10, 15] colsample_bytree: [0.8, 1.0] subsample: [0.8, 1.0]
LightGBM	learning_rate: [0.01, 0.03, 0.05, 0.1] n_estimators: [100, 200, 500] max_depth: [-1, 10, 20, 50] num_leaves: [31, 50, 100]
Ridge	alpha: [1e-15, 1e-10, 1e-8, 1e-3, 1e-2, 1, 5, 10, 20, 30, 35, 40, 45, 50, 55, 100]
Lasso	alpha: [1e-15, 1e-10, 1e-8, 1e-3, 1e-2, 1, 5, 10, 20, 30, 35, 40, 45, 50, 55, 100]
Elastic Net	alpha: [1e-15, 1e-10, 1e-8, 1e-3, 1e-2, 1, 5, 10, 20, 30, 35, 40, 45, 50, 55, 100] l1_ratio: [0.0 , 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
Linear Regression	`{}` (No hyperparameters)

Table 2: Hyperparameter Grids for Model Tuning

3.9 Model Evaluation

The performance of the models was evaluated using a comprehensive set of metrics to assess their accuracy and reliability. These metrics include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2).

- **Mean Absolute Error (MAE)**: MAE measures the average magnitude of errors in the model's predictions, providing a straightforward indication of prediction accuracy. It is defined mathematically as:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| , \tag{12}$$

  where:

  - N is the total number of observations,
  - y_i is the actual property price for the i-th observation,
  - $\hat{y}_i$ is the predicted property price for the i-th observation.

MAE provides an intuitive measure of average error in the same units as the target variable, making it easy to interpret.

- **Root Mean Squared Error (RMSE)**: RMSE expresses the error in the same units as the target variable by taking the square root of expected value of the squares of the errors or deviations. It is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \tag{13}$$

  Here, the variables N, y_i, and $\hat{y}_i$ are as defined in Equation 12. RMSE penalizes larger errors more heavily than MAE, providing insight into the distribution of prediction errors.

- **Mean Absolute Percentage Error (MAPE)**: MAPE expresses the accuracy of the model's predictions as a percentage, facilitating the comparison of prediction performance across different datasets or models. It is defined as:

$$\text{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|. \tag{14}$$

  In this equation, the variables N, y_i, and $\hat{y}_i$ are as defined in Equation 12. MAPE provides a relative measure of error, making it useful for understanding the model's performance in percentage terms.

- **Coefficient of Determination (R^2)**: R^2 evaluates how well the model explains the variance in the target variable. It is given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \tag{15}$$

  where:

  - $\bar{y}$ is the mean of the actual property prices:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i.$$

  - N is the total number of observations,
  - y_i is the actual property price for the i-th observation,
  - $\hat{y}_i$ is the predicted property price for the i-th observation.

  An R^2 score closer to 1 indicates that the model explains a higher proportion of the variance in the target variable, signifying stronger predictive power.

To ensure the robustness and generalization of the models, a 5-fold cross-validation strategy was applied within the grid search framework. This approach involves partitioning the dataset into five subsets, training the model on four subsets, and validating it on the remaining subset. This process is

repeated five times, with each subset serving as the validation set once. The cross-validation technique helps prevent overfitting by ensuring that the model's performance is consistently evaluated across different data partitions.
In summary, this systematic approach allowed for a thorough comparison of each model's ability to handle unseen data, ultimately identifying the most effective algorithm for real estate price prediction.

3.10 Conclusion

This chapter presented a comprehensive methodology for predicting real estate prices in Vilnius using machine learning. The data was collected through web scraping, followed by extensive data cleaning and preprocessing. Various machine learning models, including both traditional and ensemble methods, were applied, and hyperparameter tuning was performed to optimize model performance. Model evaluation using cross-validation and performance metrics ensured the robustness and accuracy of the predictions. The next chapter will discuss the results of the modelling process.

4 Modelling and Results

This chapter presents the core findings of the real estate price prediction study in Vilnius, showcasing how different modeling approaches perform under various feature sets. The modeling process began by establishing a baseline model using Linear Regression on apartment data only. This baseline serves as a point of reference for evaluating more complex methods. Subsequently, additional data (demographics and engineered features) were incorporated, and diverse models—including Decision Tree, Random Forest, LightGBM, XGBoost, Gradient Boosting, Ridge, Lasso, and ElasticNet—were trained, tuned, and compared. By systematically comparing model performance, we reveal the best model for achieving the most accurate predictions.
All computational experiments were conducted in a Python (Anaconda distribution) environment, leveraging several core libraries and tools:

- **NumPy** and **pandas** for data handling and manipulation
- **scikit-learn** for implementing linear and basic tree-based models
- **XGBoost** and **LightGBM** for specialized gradient boosting algorithms
- **Matplotlib** or **Seaborn** for data visualization and exploratory analysis

The data were organized into three groups:

- **Group 1**: Apartment Data Only
- **Group 2**: Apartment + Demographics
- **Group 3**: Apartment + Demographics + Newly Engineered Features

Each group was used to train and test the models following a consistent modeling pipeline and hyperparameter tuning process. The chapter also interprets the predictive performance of each approach, highlighting how the inclusion of additional features, especially newly engineered ones, can enhance model accuracy.

4.1 Data Splitting

For all modeling experiments, the dataset in each group was split into training and testing subsets. The typical split was **70%** training and **30%** testing, ensuring that models did not train on test data.

4.2 Baseline Model

To establish a point of comparison for evaluating the performance of advanced machine learning models, a Baseline Linear Regression model was implemented on the apartment data only. Linear Regression was chosen as the baseline due to its simplicity, interpretability, and widespread use in predictive modeling tasks. The baseline model provides a benchmark against which the effectiveness of more complex models can be measured.

The performance of the Baseline Linear Regression model was evaluated on the test dataset using R^2, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). These metrics offer a comprehensive view of the model's ability to capture the underlying patterns in the data and its predictive accuracy.

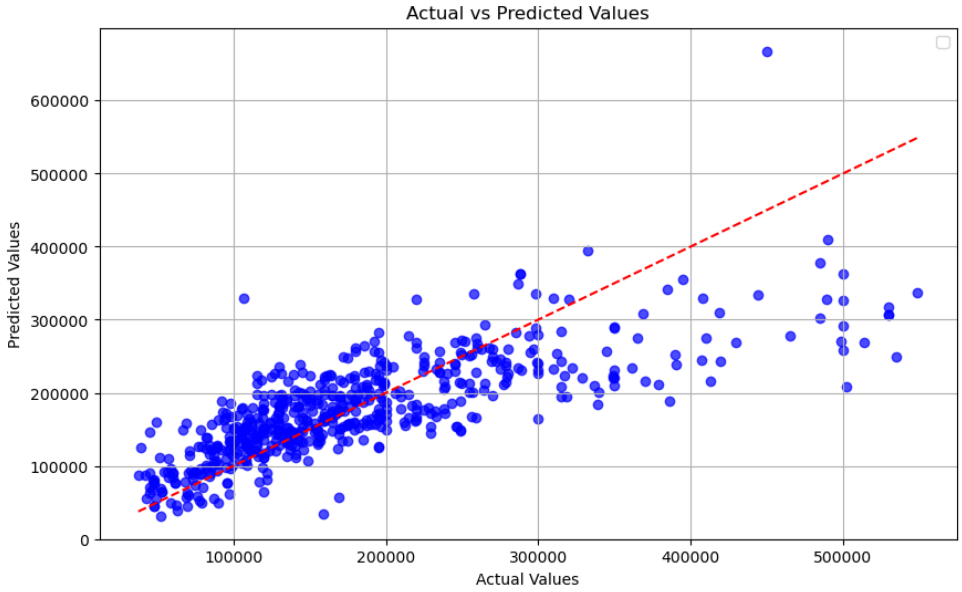


Figure 16: Baseline Model: Actual vs. Predicted Values for Linear Regression.

Metric	**Value**
R^2	0.5886
MAE	44,666.68
RMSE	63,482.78
MAPE	27.54%

Table 3: Baseline Linear Regression — Test Results (Group 1)

The results of the Baseline Linear Regression model, as presented in Table 3, indicate that while the model explains approximately 58.86% of the variance in the test dataset ($R^2 = 0.5886$), it exhibits significant prediction errors, as evidenced by the high values of MAE (44,666.68), RMSE (63,482.78), and MAPE (27.54%). These findings highlight the limitations of the linear regression model in accurately capturing the complex relationships present in the data. Similiarly, the scatter plot in Figure 16 illustrates the relationship between actual and predicted values for the baseline model. The dashed red line represents the ideal case where the predicted values are perfectly aligned with the actual values.

While the plot demonstrates that the baseline model captures the general trend of the data, several discrepancies are evident, particularly for higher values.

This baseline model serves as a foundational benchmark for comparing the performance of more advanced models, such as Random Forest, XGBoost, and other machine learning algorithms explored in this research. By improving upon the baseline metrics, the advanced models demonstrate their potential in providing more accurate and reliable predictions for the problem at hand.

4.3 Results by Data Group

The performance of each model was evaluated across three groups of data. Tables of the best results (after hyperparameter tuning) are provided, along with a concise discussion.

4.4 Group 1: Apartment Data Only

This subset includes only apartment features (e.g., floor area, number of rooms, renovation status, build year), as detailed in Table 8 in Appendix 2. .

Model	Test R^2	Test MAE	Test RMSE	Test MAPE
DecisionTree	0.6838	34,840.03	55,652.28	19.65%
RandomForest	0.8261	25,985.27	41,275.94	15.26%
LightGBM	0.8465	23,390.20	38,774.74	13.39%
XGBoost	**0.8629**	**22,189.87**	**36,651.28**	**12.45%**
GradientBoosting	0.8543	23,277.26	37,776.54	13.09%
Ridge	0.5886	44,666.68	63,482.78	27.54%
Lasso	0.5885	44,667.22	63,485.25	27.54%
ElasticNet	0.5885	44,667.22	63,485.25	27.54%
LinearRegression	0.5886	44,666.68	63,482.78	27.54%

Table 4: Model Performance (Apartment Data only)

Best Model: XGBoost achieved an R^2 of 0.863, RMSE of 36,651, MAPE of 12.45%, and MAE of 22,190.

The models were trained and tuned using only apartment-related variables, revealing a distinct hierarchy in model performance based on their complexity and ability to capture nonlinear relationships. Linear models (Ridge, Lasso, ElasticNet, and Linear Regression) exhibited the lowest coefficient of determination ($R^2 \approx 0.5886$), explaining approximately 58.86% of the variance in the target variable. The Decision Tree model improved this performance with an R^2 of 0.6838. In contrast, ensemble and boosting models significantly outperformed the simpler models, achieving R^2 values between 0.8261 (Random Forest) and 0.8629 (XGBoost), thereby accounting for 82.61% to 86.29% of the variance.

Regarding error metrics, linear models reported the highest Mean Absolute Error (MAE = 44,666.68) and Root Mean Squared Error (RMSE = 63,482.78), indicating substantial prediction errors. The Decision Tree reduced these errors to MAE = 34,840.03 and RMSE = 55,652.28. Ensemble and boosting models further minimized errors, with XGBoost achieving the lowest MAE (22,189.87) and RMSE (36,651.28). Additionally, linear models had the highest Mean Absolute Percentage Error (MAPE $\approx$ 27.54%), followed by the Decision Tree (MAPE = 19.65%). Ensemble and boosting models demonstrated superior relative accuracy, with XGBoost recording a MAPE of 12.45%.

These findings suggest that the apartment data alone contains complex, nonlinear patterns that simpler models are unable to effectively capture, thereby establishing a baseline for evaluating the impact of incorporating additional data and features.

4.5 Group 2: Apartment Data + Demographics Data

This subset includes not only apartment features but also demographic indicators such as total population, income levels, marital status distributions, and population density, as detailed in Table 9 in Appendix 2. .

Model	Test R^2	Test MAE	Test RMSE	Test MAPE
DecisionTree	0.7104	33779.16	53262.89	19.54%
RandomForest	0.8234	26321.08	41592.16	15.41%
LightGBM	0.8424	23788.72	39286.44	13.62%
XGBoost	**0.8576**	**22205.53**	**37346.74**	**12.69%**
GradientBoosting	0.8581	22901.10	37276.13	13.16%
Ridge	0.6138	43177.23	61503.66	26.86%
Lasso	0.6144	43067.84	61454.05	26.72%
ElasticNet	0.6152	43067.44	61394.62	26.75%
LinearRegression	0.6138	43177.23	61503.66	26.86%

Table 5: Model Performance (Apartment Data + Demographics Data)

Best Model: XGBoost achieved an R^2 of 0.8576, RMSE of 37,346, MAPE of 12.69%, and MAE of 22,205.

The inclusion of demographic data alongside apartment-related variables had mixed effects on model performance, with some models experiencing marginal improvements and others showing slight declines compared to the first group. Ridge, Lasso, ElasticNet, and Linear Regression demonstrated minimal changes. R^2 values increased slightly from 0.5886 to approximately 0.615, indicating a slight improvement in variance explained. However, error metrics such as MAE ($\sim 43,000$), RMSE ($\sim 61,500$), and MAPE ($\sim$ 26.8%) remained largely unchanged, highlighting the limitations of these models in capturing additional complexity introduced by demographic data. The Decision Tree model showed a slight improvement, with R^2 increasing from 0.6838 to 0.7104, accompanied by small reductions in MAE (from 34,840 to 33,779) and RMSE (from 55,652 to 53,263). MAPE showed a 0.01 decrease to 19.5%. While demographic data improved the Decision Tree's performance slightly, its predictive capability remained limited compared to more advanced models. Ensemble models like Random Forest and boosting methods such as LightGBM, XGBoost, and Gradient Boosting continued to perform best but exhibited slight declines : Random Forest: R^2 decreased slightly from 0.8261 to 0.8234, while MAE and RMSE showed minor increases. In LightGBM, R^2 dropped marginally from 0.8465 to 0.8424, with MAPE worsening from 13.39% to 13.62%. In XGBoost R^2 decreased from 0.8629 to 0.8576, with MAPE slightly worsening from 12.45% to 12.69%. However, it still achieved the best MAE (22,205) and RMSE (37,347). Gradient Boosting R^2 improved slightly from 0.8543 to 0.8581, but the differences in MAE and RMSE were negligible, and MAPE increased slightly from 13.09% to 13.16%. Although XGBoost and Gradient Boosting have very minor differences, XGBoost has a lower Mean Absolute Error (MAE), indicating that its price predictions are closer to the actual prices, on average, compared to Gradient Boosting. XGBoost also has a lower Mean Absolute Percentage Error (MAPE), showing that XGBoost performs better relative to the scale of property prices, making it more reliable for diverse price ranges. In summary, the lack of significant gains across all models implies that demographic data alone may not meaningfully enhance the predictive capabilities of these models. While the addition of

demographic data slightly improved some models, particularly the Decision Tree, most ensemble and boosting models either experienced marginal declines or maintained performance levels similar to the first group. This suggests that demographic data did not add significant predictive value to the feature set. Linear models, despite minor improvements, remained unable to capture the complexities of the data, while ensemble and boosting models continued to outperform all others, with XGBoost remaining the most reliable choice. In summary, the lack of significant gains across the all models implies that demographic data alone may not meaningfully enhance the predictive capabilities of these models.

4.6 Group 3: Apartment + Demographics + Newly Engineered Features

In addition to demographic features, this dataset incorporated newly engineered variables such as apartment age ratio, time since renovation, neighborhood renovation ratio, and other nuanced attributes reflecting building conditions and neighborhood trends, as detailed in Table 10 in Appendix 2. .

Model	**Test R^2**	**Test MAE**	**Test RMSE**	**Test MAPE**
DecisionTree	0.6411	35871.73	59287.07	19.85
RandomForest	0.8350	25482.21	40198.92	15.00
LightGBM	**0.8698**	**21926.06**	**35714.16**	**12.54**
XGBoost	0.8633	21577.75	36594.92	12.20
GradientBoosting	0.8503	22985.26	38298.49	12.80
Ridge	0.7283	36425.23	51590.75	22.93
Lasso	0.7265	36189.22	51755.46	22.84
ElasticNet	0.7265	36189.22	51755.46	22.84
LinearRegression	0.7312	36350.90	51313.43	22.97

Table 6: Model Performance (Apartment + Demographics + Newly Engineered Features)

Best Model: LightGBM achieved an R^2 of 0.8698, RMSE of 37,776, MAPE of 12.54%, and MAE of 21,926.

Including newly engineered features alongside apartment and demographic data led to significant changes in model performance, with some models showing positive improvements while others exhibited mixed results. Linear models, such as Ridge, Lasso, ElasticNet, and Linear Regression, experienced their largest improvements compared to previous groups, with R^2 increasing to $\sim$ 0.728–0.731 and reductions in MAE (to $\sim 36,000$) and RMSE (to $\sim 51,500$), although their MAPE remained high at $\sim$ 22.9%. The Decision Tree model, however, showed a decline, with R^2 dropping from 0.7104 to 0.6411, MAE increasing to 35,872, and RMSE rising to 59,287, indicating that it struggled with the complexity introduced by the engineered features. Ensemble and boosting models continued to lead in performance, with LightGBM achieving the highest R^2 (0.8698) and lowest MAE (21,926) and RMSE (35,714), along with a competitive MAPE (12.54%). XGBoost followed closely, with R^2 at 0.8633, MAE at 21,578, RMSE at 36,595, and the lowest MAPE at 12.20%. Gradient Boosting and Random Forest performed well but were slightly less accurate, with R^2 values of 0.8503 and 0.8350, respectively. These results highlight the effectiveness of engineered features in improving the predictive accuracy of advanced models like LightGBM and XGBoost, while simpler models, such as Decision Tree and linear approaches, struggled to fully leverage the additional complexity. Overall, the newly engineered features significantly enhanced the performance of ensemble methods, reaffirming their suitability for

capturing complex relationships in the data.

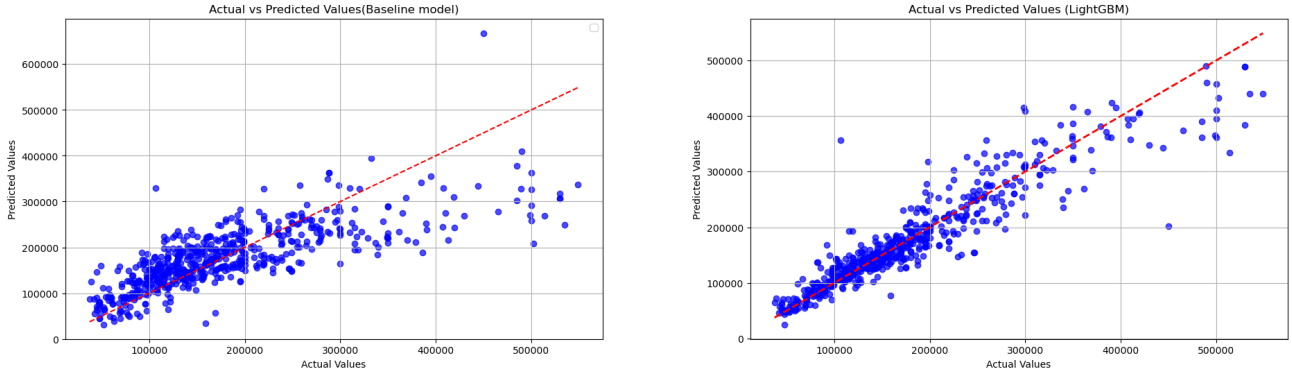
4.7 Comparative Analysis and Discussion

Table 7 summarizes the best-performing models across the three data groups, demonstrating the progression of model performance with the addition of demographic data and newly engineered features.

Group	Best Model	R^2	RMSE	MAPE	MAE
Apartment Data	XGBoost	0.8629	36,651	12.45%	22,190
Apartment + Demographics	XGBoost	0.8576	37,347	12.69%	22,205
Apartment + Demographics + Engineered Features	LightGBM	0.8698	35,714	12.54%	21,926

Table 7: Best Models Across Different Data Groups

The scatterplots in Figures 17a and 17b provide a visual comparison of prediction accuracy between the baseline model (Linear Regression) and the best-performing model (LightGBM).



(a) Linear Regression (Baseline Model)

(b) LightGBM (Best Model)

Figure 17: Comparison of Predicted vs. Actual Values for Baseline and Best Models

The comparison between the baseline Linear Regression model (built solely on apartment data) and the LightGBM model (trained on apartment, demographic, and newly engineered features) highlights the importance of feature augmentation and advanced modeling techniques. The scatterplot for Linear Regression (Figure 17a) reveals a substantial dispersion of predicted values around the ideal fit line, especially for higher-priced properties. This indicates that the baseline model struggled to capture the complexity of apartment price prediction when limited to basic apartment-related variables. In contrast, the scatterplot for LightGBM (Figure 17b) demonstrates a marked improvement, with predictions closely clustering around the ideal line. This improvement reflects the impact of incorporating demographic and engineered features, which enabled LightGBM to capture nonlinear and interaction effects in the data. Combined with superior quantitative metrics (Table 7), this comparison underscores LightGBM's ability to handle the intricate relationships between predictors and target variables, making it a robust choice for real estate pricing tasks.

4.8 Impact of Demographics

The transition from Group 1 (Apartment Only) to Group 2 (Apartment + Demographics) did not yield significant performance improvements. Although XGBoost remained the best-performing model in both groups, there was a slight decrease in R^2 (from 0.8629 to 0.8576) and marginal increases in RMSE and MAPE. These results suggest that the demographic variables included in the analysis provided limited additional predictive value, potentially due to redundancy with apartment-specific attributes or insufficient relevance to price prediction.

4.9 Effectiveness of Feature Engineering

In Group 3 (Apartment + Demographics + Newly Engineered Features), the introduction of newly engineered features led to significant improvements in model performance. The LightGBM model achieved the highest coefficient of determination ($R^2 = 0.8698$), the lowest Root Mean Squared Error (RMSE) of 35,714, Mean Absolute Error (MAE) of 21,926, and a reduction in Mean Absolute Percentage Error (MAPE) to 12.54%. These results underscore the critical role of domain-specific feature engineering in accurately capturing the complex relationships within apartment pricing data. Among the top influential features, Area and Apartment Size Ratio were the most significant, highlighting the importance of an apartment's physical size and proportional dimensions in determining its market value. Additionally, Room Density and Floor Ratio reflected the impact of room distribution and space utilization on pricing. Features related to renovation, such as Time Since Renovation and Neighborhood Renovation Ratio, emphasized the value of property upkeep and neighborhood improvements. Proximity to key amenities, including Transport Proximity Rank, Nearest Educational Institution, Nearest Kindergarten, and Nearest Shop, consistently influenced property desirability, highlighting the importance of accessibility to essential services (see Figure 18).
Furthermore, the inclusion of socioeconomic indicators, such as Marriage Rate and AVG_BIRTH_YEAR, provided additional context to the model, although their impact was relatively modest compared to structural and locational features. This observation suggests that while demographic factors contribute to the overall understanding of property valuation, the primary enhancements in model performance were driven by meticulous feature engineering focused on physical and environmental attributes.

Overall, the analysis confirms that integrating diverse features related to physical characteristics, renovation status, proximity to amenities, and selected socioeconomic indicators substantially boosts the model's predictive capability. This demonstrates the importance of comprehensive feature engineering in real estate valuation, enabling the model to effectively capture the multifaceted determinants of apartment prices.

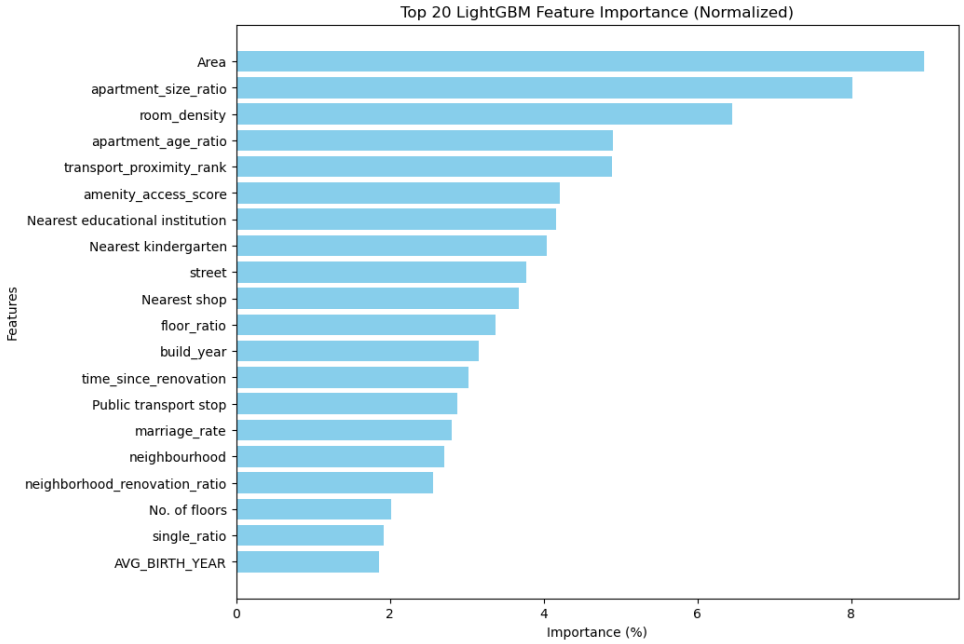


Figure 18: Top 20 LightGBM Feature Importance

4.10 Conclusion

In summary, the results demonstrate that incorporating demographic and newly engineered features significantly improves the accuracy of real estate price predictions. Among the tree-based models, **LightGBM** proved the most robust in the final data group (Group 3), achieving an R^2 of 0.8698. Overall, these findings affirm the importance of expanding beyond basic apartment characteristics to include socioeconomic context and engineered variables that capture subtle, location-specific factors.

The next chapter will synthesize these findings and outline potential avenues for future work, including how these models might be further refined or adapted to other real estate markets.

5 Conclusion

This study aimed to build predictive models for apartment prices in Vilnius using machine learning techniques. The apartment dataset was collected through web scraping one of the top real estate websites in Vilnius. To enhance the dataset, demographic data was added. Modeling was done in three groups—apartment data alone, apartment data with demographic variables, and apartment data with demographic variables and newly engineered features.

Key Findings

The baseline Linear Regression model, built exclusively on apartment data, showed limited ability to predict prices, with an R^2 of 0.5886 and significant errors. Ensemble methods like XGBoost and LightGBM performed much better. LightGBM, trained on apartment data enriched with demographics and engineered features, achieved the best results, with an R^2 of 0.8698, RMSE of 35,714, MAE of 21,926, and MAPE of 12.54%. These results highlight its ability to capture complex relationships effectively.

Adding demographic data alone resulted in marginal improvements. For instance, XGBoost's R^2 decreased slightly from 0.8629 (apartment data only) to 0.8576. This suggests that demographic variables, while relevant, added limited predictive value.

Engineered features, such as proximity to amenities and neighborhood renovation ratios, provided the most significant improvements, emphasizing the importance of well-designed features in predictive modeling.

Limitations and Future Work

Despite the promising results achieved in this study, several limitations warrant consideration. Firstly, the generalizability of the findings is constrained by the study's focus on the Vilnius real estate market. Real estate market structures and dynamics can vary substantially across different regions, implying that the developed models may require recalibration to maintain accuracy when applied to other geographical contexts. Secondly, the study was limited by data. It did not utilize the actual prices for which the apartments sold but instead relied on apartment estimates, as the data were scraped from a real estate advert site. This approach may introduce inaccuracies, potentially affecting the models' predictive performance. Furthermore, limited access to comprehensive historical real estate data posed another constraint. Historical data are essential for identifying long-term trends and seasonal patterns, and their scarcity may have restricted the model's capacity to fully understand and predict market fluctuations. Additionally, computational constraints influenced the model tuning and training processes. The tuning and training were performed on a MacBook Pro (2018) equipped with 16 GB of RAM, which restricted the complexity and scale of models that could be efficiently explored. These hardware limitations resulted in longer computational times and limited the extent of hyperparameter optimization, potentially impacting the overall performance and robustness of the models developed.

These limitations highlight areas for future research, including acquiring more comprehensive historical datasets on Vilnius municipality, utilizing more advanced computational resources, and exploring

feature engineering methodologies that can mitigate bias and enhance model generalizability. In addition, feature engineering played a pivotal role in improving the accuracy of predictive models in this study. Future research could explore including more granular and diverse features, particularly those related to neighbourhood and environmental attributes. Exploring other ensemble techniques like stacking, noted in the literature but not implemented in this study, could provide valuable insights and further refine the predictive capabilities of the models developed in this study.

References

[1] Abdul-Rahman S., Mutalibs., Zulkifley N.H, and Ibrahim, I. (2021). *Advanced Machine Learning Algorithms for House Price Prediction: Case Study in Kuala Lumpur.* *International Journal of Advanced Computer Science and Applications*, 12(12).

[2] Asres, Habtamu Bishaw. (2023). *Causes of Valuation Inaccuracy in Mortgage Lending in Ethiopia.* *International Journal of Real Estate Studies*, 17(1), 120-134.

[3] Asabere, P. K., & Huffman, F. E. (2013). *The Impact of Relative Size on Home Values.* *Appraisal Journal*, 1.

[4] Atilola Moses Idowu, Norhaya Kamarudin, Kamalahasan Achu, and Ibisola Abayomi Solomon. (2016). *A Review of Valuation Impact on Property Tax.* Journal Name, Volume(Issue), Page range. DOI: 10.11113/SH.V8N4-3.1077.

[5] Benoit, K. (2011). *Linear regression models with logarithmic transformations.* London School of Economics, 22(1), pp. 23-36.

[6] Bello, N., & Sulaiman Adepoju Adetoye, W. A. (2020). *Factors influencing rental and capital values of residential investment property.*

[7] Burgess, G., Hamilton, C., Jones, M., & Muir, K. (2017). *Multigenerational living: an opportunity for UK house builders? Final report to the NHBC Foundation-source document.* University of Cambridge.

[8] Čeh, M., Kilibarda, M., Lisec, A., and Bajat, B. (2018). *Estimating the performance of random forest versus multiple regression for predicting prices of the apartments.* *ISPRS International Journal of Geo-Information*, 7(5), 168.

[9] Cevik, S., & Naik, S. (2023). *Bubble detective: City-level analysis of house price cycles.* *IMF Working Papers*, 2023(033), 1. DOI: `https://doi.org/10.5089/9798400231537.001`.

[10] Chen, Z., Zhang, R., Chen, Z., Zheng, Y., & Zhang, S. (2023). *ScTCN-LightGBM: a hybrid learning method via transposed dimensionality-reduction convolution for loading measurement of industrial material.* *Connection Science*, 35(1). DOI: `https://doi.org/10.1080/09540091.2023.2278275`.

[11] Chau, K.W. and Chin, T.L. (2003). *A critical review of literature on the hedonic price model.* *International Journal for Housing Science and its applications*, 27(2), pp.145-165.

[12] Chigwenya, A., & Dube, D. (2019). *Infrastructure development and property values in low income residential properties in Bulawayo.* *International Journal of Built Environment and Scientific Research*, 2(2), 131. DOI: `https://doi.org/10.24853/ijbesr.2.2.131-140`.

[13] Ciarlone, A. (2015). *House price cycles in emerging economies.* *Studies in Economics and Finance*, 32(1), 17–52. DOI: `https://doi.org/10.1108/sef-11-2013-0170`.

[14] Collier, P., Glaeser, E., Venables, T., Manwaring, P., & Blake, M. (2017). *Land and property taxes: exploiting untapped municipal revenues*. Policy Brief.

[15] Cohen, V., & Karpavičiūtė, L. (2017). *The analysis of the determinants of housing prices*. *Independent Journal of Management & Production*, 8(1), 49–63. DOI: https://doi.org/10.14807/ijmp.v8i1.521.

[16] Croom, B., Kennedy, S., Ojha, S., and Sparks, J. (2020). *Analysis of the commercial real estate market in a post COVID-19 world*. *SMU Data Science Review*, 3(3). Available at: https://scholar.smu.edu/datasciencereview/vol3/iss3/5.

[17] Cupal, M. (2015). *Historical perspective of residential development and its impact on the current market prices of apartments on the Czech real estate market*. *Procedia Economics and Finance*, 26, 144–151. DOI: https://doi.org/10.1016/s2212-5671(15)00902-8.

[18] Ehrlich. (2023). *The Impact Of Unemployment And The Economy On The Real Estate Market*. Available at: https://www.s-ehrlich.com/the-impact-of-unemployment-and-the-economy-on-the-real-estate-market/.

[19] Elliott, P., Miller, S., Pawar, S., Vaccaro, B. J., McCullough, M., Rao, P., Ghosh, R., Warier, P., Desai, N. R., & Ahmad, T. (2019). *Predictive Abilities of Machine Learning Techniques May Be Limited by Dataset Characteristics: Insights From the UNOS Database*. *Journal of Cardiac Failure*. DOI: https://doi.org/10.1016/J.CARDFAIL.2019.01.018.

[20] Enwere, K. P., & Ogoke, U. P. (2023). *A Comparative Approach on Bridge and Elastic Net Regressions*. *African Journal of Mathematics and Statistics Studies*, 6(2), 67–79. DOI: https://doi.org/10.52589/AJMSS-LBJO9UCU.

[21] Francke, M., & Korevaar, M. (2020). *Baby Booms and Asset Booms: Demographic Change and the Housing Market*. *Social Science Research Network*. DOI: https://doi.org/10.2139/SSRN.3368036.

[22] Gao, Q., Shi, V., Pettit, C., & Han, H. (2022). *Property valuation using machine learning algorithms on statistical areas in Greater Sydney, Australia*. *Land Use Policy*, 123. DOI: 10.1016/j.landusepol.2022.106409.

[23] Gasparėnienė, L., Remeikienė, R., and Skuka, A. (2016). *Assessment of the impact of macroeconomic factors on housing price level: Lithuanian case*. *Intellectual Economics*, 10(2), 122–127. DOI: https://doi.org/10.1016/j.intele.2017.03.005.

[24] Geng, N. (2018). *Fundamental Drivers of House Prices in Advanced Economies*. *IMF Working Papers*, 18(1). DOI: 10.5089/9781484367629.001.

[25] Gevorgyan, K. (2019). *Do demographic changes affect house prices*. *Cambridge University Press*, 85(4), 305–320. DOI: https://doi.org/10.1017/DEM.2019.9.

[26] Glaeser, E., & Gyourko, J. (2018). *The economic implications of housing supply.* *The Journal of Economic Perspectives: A Journal of the American Economic Association*, 32(1), 3–30. DOI: https://doi.org/10.1257/jep.32.1.3.

[27] Gong, Y., & Yao, Y. (2022). *Demographic changes and the housing market.* *Regional Science and Urban Economics*, 95(103734), 103734. DOI: https://doi.org/10.1016/j.regsciurbeco.2021.103734.

[28] Grybauskas, A., Pilinkiene, V., & Stundziene, A. (2021). *Predictive Analytics Using Big Data for the Real Estate Market During the COVID-19 Pandemic.* *Journal of Big Data*, 8(1), 105. DOI: 10.1186/s40537-021-00522-z.

[29] Harvard Law Review. (2022). *Addressing Challenges to Affordable Housing in Land Use Law: Recognizing Affordable Housing as a Right.* *Harvard Law Review*, 4(135), 1104–1125. Available at: https://harvardlawreview.org/print/vol-135/addressing-challenges-to-affordable-housing-in-land-use-law/.

[30] Hoffmann, J.P. (2021). *Linear regression models: applications in R.* *Chapman and Hall/CRC*.

[31] Ho, W.K., Tang, B.S., and Wong, S.W. (2021). *Predicting property prices with machine learning algorithms.* *Journal of Property Research*, 38(1), 48-70.

[32] Huang, Y., & Dall'erba, S. (2021). *Does proximity to school still matter once access to your preferred school zone has already been secured?.* *Journal of Real Estate Finance and Economics*, 62(4), 548–577. DOI: https://doi.org/10.1007/s11146-020-09761-w.

[33] Kendall, R., & Tulip, P. (2018–2021). *The effect of zoning on housing prices.* *Reserve Bank of Australia Research Discussion Paper*.

[34] Kok, N., Monkkonen, P., & Quigley, J. M. (2014). *Land use regulations and the value of land and housing: An intra-metropolitan analysis.* *Journal of Urban Economics*, 81, 136–148. DOI: https://doi.org/10.1016/j.jue.2014.03.004.

[35] Kouki, T. (2018). *The Effects of Government Policies on Real Estate Sector.* *KTH Royal Institute of Technology*. Available at: https://urn.kb.se/resolve?urn=urn:nbn:se:kth.

[36] Krajnakova, E., Jegelavičiūtė, R., and Navickas, V. (2018). *The economic factors influence on real estate market development.* *Ad Alta: Journal of Interdisciplinary Research*, 8, 141–146.

[37] Krolage, C. (2023). *The effect of real estate purchase subsidies on property prices.* *International Tax and Public Finance*, 30(1), 215–246. DOI: https://doi.org/10.1007/s10797-022-09726-0.

[38] Lennon H. T. Choy and Winky K. O. Ho. *The Use of Machine Learning in Real Estate Research.* Land, 12(4), 2023, Article 740. DOI: 10.3390/land12040740. Available at: https://www.mdpi.com/2073-445X/12/4/740.

[39] Lima, R. C. de A., & Silveira Neto, R. da M. (2019). *Zoning ordinances and the housing market in developing countries: Evidence from Brazilian municipalities.* *Journal of Housing Economics*, 46(101653), 101653. DOI: https://doi.org/10.1016/j.jhe.2019.101653.

[40] Mayank, S., Rahul, C., Swati, D., Kanegonda, R., & Chythanya, R. (2024). *House Price Prediction Using Linear and Lasso Regression.* DOI: https://doi.org/10.1109/inocon60754.2024.10511592.

[41] Maha, Shabbir., Sohail, Chand., Farhat, Iqbal. (2023). *A new ridge estimator for linear regression model with some challenging behavior of error term.* *Communications in Statistics - Simulation and Computation*. DOI: https://doi.org/10.1080/03610918.2023.2186874.

[42] Mindi, R., Putri, I. G. P. S., Wijaya, F. P., Praja, A. H., Hadi, A., & Hamami, F. (2023). *The Comparison Study of Regression Models (Multiple Linear Regression, Ridge, Lasso, Random Forest, and Polynomial Regression) for House Price Prediction in West Nusa Tenggara.* DOI: https://doi.org/10.1109/icadeis58666.2023.10270916.

[43] Monson, M. (2009). *Valuation using hedonic pricing models.*

[44] Muhammed, U., Sani, I., Doguwa, B. B., & Alhaji. (2022). *Comparing the Prediction Accuracy of Ridge, Lasso and Elastic Net Regression Models with Linear Regression Using Breast Cancer Data.* *Bayero Journal of Pure and Applied Sciences*. DOI: https://doi.org/10.4314/bajopas.v14i2.16.

[45] Natekin, A., & Knoll, A. (2013). *Gradient boosting machines, a tutorial.* *Frontiers in Neurorobotics*, 7, 21.

[46] Naz, Rabia, Jamil, Bushra, and Ijaz, Humaira. (2024). *Real Estate Price Prediction.* *International Journal of Information Science and Technology*, 6, 1031-1044.

[47] Nazarov, F. M. (2023). *Optimization of Prediction Results Based on Ensemble Methods of Machine Learning.* DOI: https://doi.org/10.1109/SmartIndustryCon57312.2023.10110726.

[48] Norchi, F. M. *Affordable Housing Development Toolkit: Successful approaches from North Carolina and beyond*, 2019.

[49] Nurudeen Akinsola Bello and Olusegun Olaopin Olanrele. (2016). *Value gap in Nigerian property compensation practice: measurement and economic effects.* *Pacific Rim Property Research Journal*, 22(2), 101-113. DOI: 10.1080/14445921.2016.1203235.

[50] OECD. (2023). *The Lithuanian housing market: Quality and affordability gaps in a challenging policy context.* *OECD iLibrary*. Available at: https://www.oecd-ilibrary.org/sites/d4ec4569-en/index.html?itemId=/content/component/d4ec4569-en.

[51] Okonkwo, I. (2023). *Understanding real estate appraisal and valuation.* *Business-Day Media Limited*. Available at: https://businessday.ng/life-arts/article/understanding-real-estate-appraisal-and-valuation/.

[52] Oktay, E., Karaaslan, A., Alkan, Ö., and Kemal Çelik, A. (2014). *Determinants of housing demand in the Erzurum province, Turkey.* *International Journal of Housing Markets and Analysis*, 7(4), 586–602. DOI: https://doi.org/10.1108/ijhma-11-2013-0056.

[53] Obondy, S. (2013). *The effect of interest rates on the supply of real estate finance in Nairobi County.*

[54] Owusu-Ansah, A. (2011). *A review of hedonic pricing models in housing research.* *Journal of International Real Estate and Construction Studies*, 1(1), 19.

[55] Park, B., & Bae, J. K. (2015). *Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data.* *Expert Systems with Applications*, 42(6), 2928-2934.

[56] Pang, F. (2023). *Selling Your Investment Property to First-Time Home Buyers: A comprehensive guide.* *Com.au*. Available at: https://www.boldre.com.au/post?post_id=12488.

[57] Post, J. E., and Berkhout, T. (2014). *Risk perceptions in the European real estate industry.* Retrieved February 24.

[58] Rigatti, S. J. (2017). *Random forest.* *Journal of Insurance Medicine*, 47(1), 31-39.

[59] Rivas, R., Patil, D., Hristidis, V., Barr, J. R., & Srinivasan, N. (2019). *The impact of colleges and hospitals to local real estate markets.* *Journal of Big Data*, 6(1). DOI: https://doi.org/10.1186/s40537-019-0174-7.

[60] Rutherford, J., Rutherford, R. C., Strom, E., & Wedge, L. (2017). *The subsequent market value of former REO properties.* *Real Estate Economics*, 45(3), 713–760. DOI: https://doi.org/10.1111/1540-6229.12134.

[61] Shumway, R. H., Stoffer, D. S., Shumway, R. H., & Stoffer, D. S. (2017). *ARIMA models. Time Series Analysis and Its Applications: With R Examples* (pp. 75-163).

[62] Sibindi, R., Mwangi, R. W., & Waititu, A. G. (2022). *A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices.* *Engineering Reports*. DOI: https://doi.org/10.1002/eng2.12599.

[63] Siroya, S. (2023). *Residential property: Importance of location when buying a house.* *Times of India*, March 4. Available at: https://timesofindia.indiatimes.com/bs/voices/residential-property-importance-of-location-when-buying-a-house/.

[64] Song, Y. Y., & Ying, L. U. (2015). *Decision tree methods: applications for classification and prediction.* *Shanghai Archives of Psychiatry*, 27(2), 130.

[65] Squires, G., & White, I. (2019). *Resilience and housing markets: Who is it really for?.* *Land Use Policy*, 81, 167–174. DOI: https://doi.org/10.1016/j.landusepol.2018.10.018.

[66] Sturtevant, L. (2018). *The impacts of rent control: A research review and synthesis.* *National Multifamily Housing Council*.

[67] Takáts, E. (2012). *Aging and house prices.* *Journal of Housing Economics*, 21(2), 131–141. DOI: https://doi.org/10.1016/j.jhe.2012.04.001.

[68] Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). *Housing price prediction via improved machine learning techniques.* *Procedia Computer Science*, 174, 433-442.

[69] Tyvimaa, T., & Kamruzzaman, M. (2019). *The effect of young, single person households on apartment prices: an instrument variable approach.* *Journal of Housing and the Built Environment*, 34(1), 91–109. DOI: https://doi.org/10.1007/s10901-018-9618-1.

[70] Umar, I. (2021). *Impact of Infrastructural Facilities on Residential Property Rental Value.* *Abubakar Tafawa Balewa University*. Available at: https://www.researchgate.net/publication/353259705_Impact_of_Infrastructural_Facilities_on_Residential_Property_Rental_Value.

[71] Vetter, D. M., Beltrao, K. I., & Massena, R. (2013). *The impact of the sense of security from crime on residential property values in Brazilian metropolitan areas.* *SSRN Electronic Journal*. DOI: https://doi.org/10.2139/ssrn.2367688.

[72] Wang, Kai. (2023). *Predicting Real Estate Price Using Stacking-Based Ensemble Learning.* *American Journal of Information Science and Technology*. DOI: https://doi.org/10.11648/j.ajist.20230702.14.

[73] Ye, Qiongwei. (2024). *House price prediction using machine learning for Ames, Iowa.* *Applied and Computational Engineering*, 55, 44-54. DOI: https://doi.org/10.54254/2755-2721/55/20241483.

[74] Zahirovich-Herbert, V., & Gibler, K. M. (2014). *The effect of new residential construction on housing prices.* *Journal of Housing Economics*, 26, 1–18. DOI: https://doi.org/10.1016/j.jhe.2014.06.003.

[75] Zekun Chen. (2024). *Integrating Machine Learning Techniques for Real Estate Analysis.* *Highlights in Science Engineering and Technology*, 92, 252-256. DOI: https://doi.org/10.54097/cvn9mr84.

[76] Zheng, Y. (2017). *The impact of population age structure on real estate price—evidence from China provincial panel data.* *Open Journal of Social Sciences*, 05(03), 212–222. DOI: https://doi.org/10.4236/jss.2017.53019.

[77] Zozaya, Andrea, Ramirez-Lechuga, Sharon, and Luna López, Eri. (2023). *Real Estate Price Prediction Using Regression Techniques.*

Appendix 1. Table showing distribution of building type by neighbourhood with percentages

neighbourhood	Block house	Brick	Carcass house	Log house	Monolithic	Other	Wooden house	Total	Brick%	Block%
Karoliniškės	83	20	1	0	4	1	0	109	18.348624	76.146789
Lazdynai	64	19	0	0	6	0	0	89	21.348315	71.910112
Viršuliškės	34	15	0	0	15	1	0	65	23.076923	52.307692
Fabijoniškės	69	38	1	0	9	0	0	117	32.478632	58.974359
Šeškinė	53	35	0	0	5	0	0	93	37.634409	56.989247
Verkiai	9	8	0	0	3	0	0	20	40.000000	45.000000
Žirmūnai	81	67	0	0	13	2	1	164	40.853659	49.390244
Justiniškės	41	48	0	0	4	1	0	94	51.063830	43.617021
Pilaitė	53	114	0	0	12	4	0	183	62.295082	28.961749
Rasos	0	16	0	0	7	0	2	25	64.000000	0.000000
Pašilaičiai	44	114	0	0	10	1	0	169	67.455621	26.035503
Markučiai	6	42	0	0	8	1	2	59	71.186441	10.169492
Žemieji Paneriai	4	14	0	0	1	0	0	19	73.684211	21.052632
Šnipiškės	23	145	0	2	14	1	9	194	74.742268	11.855670
Žvėrynas	1	83	0	0	19	0	7	110	75.454545	0.909091
Vilkpėdė	3	19	0	0	1	2	0	25	76.000000	12.000000
Naujamiestis	56	264	0	0	19	2	0	341	77.419355	16.422287
Baltupiai	12	62	0	0	4	0	0	78	79.487179	15.384615
Bajorai	4	32	0	0	4	0	0	40	80.000000	10.000000
Naujoji Vilnia	9	65	0	1	1	0	5	81	80.246914	11.111111
Antakalnis	22	154	0	1	2	0	1	180	85.555556	12.222222
Burbiškės	2	24	0	0	2	0	0	28	85.714286	7.142857
Šiaurės miestelis	1	29	0	0	0	3	0	33	87.878788	3.030303
Naujininkai	3	100	0	0	9	0	1	113	88.495575	2.654867
Lazdynėliai	4	80	0	0	1	0	0	85	94.117647	4.705882
Senamiestis	5	249	0	2	4	1	0	261	95.402299	1.915709
Paupys	0	24	0	0	1	0	0	25	96.000000	0.000000
Jeruzalė	1	43	0	0	0	0	0	44	97.727273	2.272727
Užupis	0	46	0	0	0	0	1	47	97.872340	0.000000
Santariškės	0	42	0	0	0	0	0	42	100.000000	0.000000

Figure 19: Distribution of building type by neighbourhood with percentages

Appendix 2. Feature Groups

.

The following tables provide an overview of the features used for each group in the analysis.

Group 1: Apartment Data Only

Features
Neighbourhood, Street, Area, Number of rooms, Floor, Build year, No. of floors, Building type, Equipment, No of Additional Equipments, Nearest kindergarten, Nearest educational institution, Nearest shop, Public transport stop.

Table 8: Group 1 Features: Apartment Data only

Group 2: Apartment + Demographics

Features
Neighbourhood, Street, Area, Number of rooms, Floor, Build year, No. of floors, Building type, Equipment, No of Additional Equipments, Nearest kindergarten, Nearest educational institution, Nearest shop, Public transport stop, AVG_BIRTH_YEAR, TOTAL_CHILDREN, TOTAL_PEOPLE, TOTAL_DIVORCED, TOTAL_WIDOWED, TOTAL_MARRIED, TOTAL_SINGLE.

Table 9: Group 2 Features: Apartment + Demographics

Group 3: Apartment + Demographics + Newly Engineered Features

Features
Neighbourhood, Street, Area, Number of rooms, Floor, Build year, No. of floors, Building type, Equipment, Nearest kindergarten, Nearest educational institution, Nearest shop, Public transport stop, AVG_BIRTH_YEAR, TOTAL_CHILDREN, TOTAL_PEOPLE, TOTAL_DIVORCED, No of Additional Equipments, TOTAL_WIDOWED, TOTAL_MARRIED, TOTAL_SINGLE, Has_Balcony, Has_Terrace, Has_Parking_space, Has_Sauna, Has_Attic, Has_Storeroom, Has_Cellar, Has_Closet, Is_first_floor, Is_renovated, Is_price_decreased, Building age, Time since renovation, Is_recently_renovated, Floor ratio, Low rise, Mid rise, High rise, Marriage rate, Children ratio, Single ratio, Divorced or widowed rate, Avg age, Amenity access score, Is_family_friendly, Is_retirement_area, Floor accessibility impact, Transport proximity rank, Neighborhood age diversity, Neighborhood renovation ratio, Children to family ratio, Room density, Apartment size ratio, Apartment age ratio, Floor rank, Is_new_or_renovated, Is_close_to_transport.

Table 10: Group 3 Features: Apartment + Demographics + ewly Engineered Features

Appendix 3. Newly Engineered features

Feature Name	**Formulation**
Building Age	`current_year - build_year`

Time Since Renovation	`current_year - renovation_year`
	`if renovation_year is empty,`
	`then current_year - build_year`

Is Recently Renovated	`time_since_renovation < 10`
Floor Ratio	`Floor / No. of floors`
Low Rise	`No. of floors <= 3`
Mid Rise	`(No. of floors > 3) && (No. of floors <= 9)`
High Rise	`No. of floors > 9`
Marriage Rate	`TOTAL_MARRIED / TOTAL_PEOPLE`
Children Ratio	`TOTAL_CHILDREN / TOTAL_PEOPLE`
Single Ratio	`TOTAL_SINGLE / TOTAL_PEOPLE`
Divorced or Widowed Rate	`(TOTAL_DIVORCED + TOTAL_WIDOWED) / TOTAL_PEOPLE`
Average Age	`current_year - AVG_BIRTH_YEAR`

Amenity Access Score	`1 / (Nearest kindergarten +`
	`Nearest educational institution +`
	`Nearest shop + Public transport stop)`

Is Family Friendly	`children_ratio > 0.3`
Is Retirement Area	`avg_age > 60`
Floor Accessibility Impact	`(building_age > 50) && (Floor > 3)`
Transport Proximity Rank	`rank within 'neighbourhood' based on 'Public transport stop'`
Neighborhood Age Diversity	`std(avg_age) grouped by 'neighbourhood'`
Neighborhood Renovation Ratio	`mean(is_recently_renovated) grouped by 'neighbourhood'`
Children to Family Ratio	`TOTAL_CHILDREN / (TOTAL_MARRIED + TOTAL_DIVORCED + TOTAL_WIDOWED)`
Room Density	`Number of rooms / Area`
Apartment Size Ratio	`Area / neighborhood mean Area`
Apartment Age Ratio	`building_age / neighborhood mean building_age`
Floor Rank	`rank of Floor within 'neighbourhood'`
Rooms Per Floor	`Number of rooms / Floor`
Is New or Renovated	`building_age < 5 OR is_recently_renovated`
Is Close to Transport	`Public transport stop <= 500`

Feature Name	**Formulation**
Has Balcony	`'Additional premises' contains 'Balcony'`
Has Terrace	`'Additional premises' contains 'Terrace'`
Has Parking Space	`'Additional premises' contains 'Parking space'`
Has Sauna	`'Additional premises' contains 'Sauna'`
Has Attic	`'Additional premises' contains 'Attic'`
Has Storeroom	`'Additional premises' contains 'Storeroom'`
Has Cellar	`'Additional premises' contains 'Cellar'`
Has Closet	`'Additional premises' contains 'Closet'`
Is First Floor	`Floor == 1`

Table 11: Engineered Features and Their Calculations

Appendix 4. Model Hyperparameters

Model	Group	Best Parameters
DecisionTree	Group 1	{'criterion': 'absolute_error', 'max_depth': 10, 'max_features': None, 'min_samples_leaf': 7, 'min_samples_split': 2, 'random_state': 42}
RandomForest	Group 1	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 500, 'random_state': 42}
LightGBM	Group 1	{'learning_rate': 0.05, 'max_depth': 7, 'n_estimators': 500, 'num_leaves': 31, 'random_state': 42}
XGBoost	Group 1	{'colsample_bytree': 0.8, 'learning_rate': 0.05, 'max_depth': 6, 'n_estimators': 500, 'random_state': 42, 'subsample': 0.8}
GradientBoosting	Group 1	{'learning_rate': 0.05, 'max_depth': 6, 'min_samples_split': 10, 'n_estimators': 500, 'random_state': 42}
Ridge	Group 1	{'alpha': 1e-15, 'random_state': 42}
Lasso	Group 1	{'alpha': 10, 'random_state': 42}
ElasticNet	Group 1	{'alpha': 10, 'l1_ratio': 1.0}
LinearRegression	Group 1	N/A
DecisionTree	Group 2	{'criterion': 'absolute_error', 'max_depth': None, 'max_features': None, 'min_samples_leaf': 9, 'min_samples_split': 2, 'random_state': 42}
RandomForest	Group 2	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 500, 'random_state': 42}
LightGBM	Group 2	{'learning_rate': 0.05, 'max_depth': 7, 'n_estimators': 500, 'num_leaves': 31, 'random_state': 42}
XGBoost	Group 2	{'colsample_bytree': 0.8, 'learning_rate': 0.05, 'max_depth': 6, 'n_estimators': 500, 'random_state': 42, 'subsample': 0.8}
GradientBoosting	Group 2	{'learning_rate': 0.05, 'max_depth': 6, 'min_samples_split': 2, 'n_estimators': 500, 'random_state': 42}
Ridge	Group 2	{'alpha': 1e-15, 'random_state': 42}
Lasso	Group 2	{'alpha': 100, 'random_state': 42}
ElasticNet	Group 2	{'alpha': 0.01, 'l1_ratio': 0.8}
LinearRegression	Group 2	N/A

Model	Group	Best Parameters
DecisionTree	Group 3	{'criterion': 'absolute_error', 'max_depth': 10, 'max_features': None, 'min_samples_leaf': 5, 'min_samples_split': 2, 'random_state': 42}
RandomForest	Group 3	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200, 'random_state': 42}
LightGBM	Group 3	{'learning_rate': 0.05, 'max_depth': -1, 'n_estimators': 500, 'num_leaves': 31, 'random_state': 42}
XGBoost	Group 3	{'colsample_bytree': 0.8, 'learning_rate': 0.05, 'max_depth': 6, 'n_estimators': 500, 'random_state': 42, 'subsample': 0.8}
GradientBoosting	Group 3	{'learning_rate': 0.05, 'max_depth': 6, 'min_samples_split': 10, 'n_estimators': 500, 'random_state': 42}
Ridge	Group 3	{'alpha': 0.01, 'random_state': 42}
Lasso	Group 3	{'alpha': 20, 'random_state': 42}
ElasticNet	Group 3	{'alpha': 20, 'l1_ratio': 1.0}
LinearRegression	Group 3	N/A

Table 12: Best Hyperparameters for Models Across Data Groups

Appendix 5. The use of AI tools

- **Grammarly:** Grammarly was employed to identify and correct grammar errors, and improve sentence structure.
- **ChatGPT:** ChatGPT was utilized as a coding assistant for debugging errors, and optimizing scripts.

Appendix 6. Code Repository

All code and dataset for this project are publicly available on GitHub at https://github.com/Sherlocked-Blaire/vilnius-apartment-price-prediction/tree/main.