Santeri Sjöblom

Machine Learning for Predicting the Prices of Dwellings in Small and Large Cities of Finland

Master’s Thesis in Information Systems
Supervisors: Dr. Xialou Wang
Faculty of Social Sciences, Business and Economics, and Law
Åbo Akademi University
Turku 2023

ABSTRACT

Subject: Information Systems

Writer: Santeri Sjöblom

Title: Machine Learning for Predicting the Prices of Dwellings in Small and Large Cities of Finland

Supervisor: Dr. Xiaolu Wang

Abstract:

Dwellings are one of the most expensive purchases individuals make in their lifetime. Additionally, dwellings can be considered as investment assets. Therefore, it is important to be able to precisely appraise the value of a dwelling, a task which can be achieved through the use of machine learning techniques. The first part of the study is a literature review which covers dwelling market dynamics, pricing of dwellings and use of machine learning in the field. The second part of the thesis presents a study on the Finnish dwelling markets, with a focus on the development of machine learning models for predicting dwelling prices in both large and small Finnish cities. Cities that have over 100,000 residents are considered as large cities and less than 100,000 residents small cities. The research datasets are divided based on the size of the cities into three datasets, with one containing all observations, one containing observations only from large cities, and one containing observations from small cities. The study tests different machine learning algorithms with each dataset and compares the best performing models of each dataset. The results show that the XGBoost algorithm is the best performing algorithm for predicting dwelling prices in Finnish cities. Furthermore, the study found that the importance of residents having a master’s degree in a district decreases in small cities, while it is the most important feature in large cities.

Keywords: Machine learning, Linear regression, Decision tree, Random Forest, XGBoost, ExtraTreesRegressor, Predictive analytics, Dwelling price prediction, Finnish dwelling markets

Date: 16.4.2023

Number of pages: 107

TABLE OF CONTENTS

1 INTRODUCTION

The dwelling market is a critical sector of any economy, impacting not only homeowners, but also renters, landlords, and property investors. Dwellings are valuable assets. Typically, a home is the most valuable asset one purchases during one's life and housing costs comprise a great share of an individual's income. Dwellings can be considered as a type of investments too and as a result, individuals want, or at least should want, to know whether they are using their money wisely when buying a dwelling and whether they are offering the correct price to a seller or not. One of the main challenges in the dwelling market is correctly predicting dwelling prices, which are influenced by multiple factors, such as dwelling characteristics, districts' demographics, location, and market trends.

In Finland, the value of the dwelling market was 13.8 billion euros in 2022 (FREA 2023), thus the dwelling market is significant. At the end of 2020, Finland had 3,124,000 dwellings, 358,000 of them without a permanent resident. The average dwelling's floor area was 79.4 square meters. Finns live in different type of dwellings since 47% of the Finnish residential buildings were apartments, while the number of one-family houses was 1,179,000 and that of terraced houses was 420,000 (OSF, 2021).

1.1 Machine Learning

Machine learning can solve a vast number of different problems. Machine learning is an assortment of different algorithms and techniques, which are used to create a new system from data. Traditionally, users gave data and a program to a computer, which created an output based on these. Machine learning changes that, while users give data and an output to a computer, which learns from the given information by using algorithms, and creates a program (Wei-Meng, 2019). With the advent of machine learning, the ability to predict dwelling prices has considerably enhanced, therefore machines can be a major help for different dwelling market stake holders.

1.2 Objective and Research Questions

Regional differences, such as economic and demographic differences, typically affect dwelling prices (Subasi 2020). Thus, the objective of this thesis is to develop multiple machine learning models capable of predicting the price of an apartment or a real estate in the Finnish dwelling market when provided with certain input features. The models will be trained using data gathered from Finnish cities. Some of the models will be trained by data collected from cities, which have fewer than 100,000 residents and some by using data from Finnish cities that have more than 100,000 residents. This study will assess and discuss the quality of the developed models, compare them, and try to find the best performing ones for cities with different population sizes. In addition, the study will try to find the most important features of the best performing models. Furthermore, the author will discuss the discrepancies of the models developed with different datasets and try to understand the reasons behind them. The motivation for this research comes from the author's occupation which is related to dwellings and their valuations.

Large cities have more inhabitants and typically a more diverse and wider range of dwelling options in terms of dwelling features but also prices, due to their larger population. In contrast, the dwelling market in smaller cities might be more tied to the local economy and the available dwelling options are usually restricted. Therefore, the characteristics of the two datasets might vary, i.e., the dataset of the large cities is more complex than the dataset of the small cities. This suggests that dwelling price prediction models may need to be tailored to the specific characteristics of each dataset, depending on the size of the city.

To conduct this thesis, data were collected from asuntojen.hintatiedot.fi, a database maintained by the Finnish Federation of Real Estate Agencies. Additionally, demographic and socioeconomic information for each micro-location were gathered from the Paavo database, which is maintained by Statistics Finland. The data gathered from Asuntojen.hintatiedot.fi provide comprehensive information on real dwelling transaction prices and features, such as condition, type, location, and size. The Paavo database contains information on Finnish postal code areas, such as population structure, income and education level, building structure, and workplaces.

The collected data were cleaned and pre-processed to make it suitable for model creation. The pre-processed data were then used to train the models which created patterns based on the input data and used them to predict dwelling prices. The author developed the models by using the Python's Scikit-learn library, which allowed the use of well-known machine learning algorithms easily.

The following research questions are addressed in the thesis to find answers to the objectives mentioned:

1. How can machine learning be used to predict the price of a residential building in Finland?
2. Which of the developed machine learning models is the most reliable in predicting the price of a dwelling in a specific area?
3. Which features of a dataset are the most significant in predicting the price of a dwelling?
4. Which are the most significant discrepancies between the models developed in different areas?

1.3 Methodology

Various methods are used to gather necessary understanding to answer the research questions. The first question is answered by using knowledge gathered in Chapters 2, 3, and 4. The information in Chapter 2 and 3 is gained by literature review. The chapters cover overall dwelling markets and dwelling markets in Finland. Furthermore, the chapters discuss the usage of machine learning methods and machine learning opportunities in the real estate business. The author will answer the second research question based on the literature review in Chapter 3 and the empirical study conducted in Chapter 4. Furthermore, the empirical part is used to answer the research questions 3 and 4. The machine learning models developed for this study will be compared and the study will discuss their relevance and differences and try to find the reasons behind the models' performance. This study focuses on the dwellings sales market, not on the rental market.

Previous studies regarding the price prediction of Finnish individual dwellings by machine learning concentrate on Helsinki metropolitan area (Kalliola et al., 2021; Laaksonen, 2022). Oikarinen, Engblom, (2015) and Dufitinema (2020) examined models in different Finnish cities, but they concentrate on dwelling price indexes, not individual dwellings. This thesis distinguishes from those studies by using actual transaction data, not asking price data. In addition, this study concentrates on the prices of individual dwellings in different Finnish cities. The cities are different sized, and they have other different characteristics, for instance, some cities have a university while some does not. Furthermore, the city of Helsinki is excluded due to its important role in prior studies, even though it is the largest city of Finland.

1.4 The Structure

This thesis is comprised of five chapters. The introduction chapter represents information about the structure of the thesis, a short introduction on the Finnish dwelling market and machine learning and establish the objectives of this thesis. Additionally, the introduction chapter explains the methods used in this thesis.

The second and third chapters will familiarize the reader more deeply with the dwelling market and machine learning, which is crucial in comprehending the objectives and results of this thesis. The third chapter elaborates on different machine learning concepts, for instance, supervised and unsupervised learning, and the steps of a machine learning model building. In addition, the chapter explains how machine learning works in the dwelling market and how it can be used in order to make reliable predictions.

The fourth chapter is for the empirical study which utilizes the knowledge gained in the prior chapters. The first part of the empirical study discusses about the data gathered and how they have been processed in order to make the data efficient for a model building. Next, the model building process is explained. Finally, the author analyzes the results of the models. The final chapter is for the conclusion of the thesis project.

2 DWELLING MARKETS

The dwelling market is an essential component of any economy, impacting a wide range of individuals and businesses, from homeowners and renters to real estate agents and property investors. Understanding the dynamics of the housing market is critical for making informed decisions and predicting future trends in the market.

In this chapter, the overall global trends in the housing market are explored, with a particular focus on the Finnish dwelling market. The key drivers of the dwelling market are examined, including demographics, economic conditions, and government policies, and how they impact dwelling prices and demand.

After the feudal systems that were dominant in medieval Europe, private landownership has risen ever since. The Industrial Revolution and innovations in the agriculture caused a proliferation of population. Increase of productivity in agriculture meant that people were not needed in the fields anymore to the same extent, hence they started to seek other tasks, usually in urban centers. The migration wave to cities generated a great need for dwellings (Ryan-Collins et al. 2017).

For most, dwellings represent homes. On the one hand, they are places, where individuals can go from work, spend time with a family and sleep in a shelter. On the other hand, dwellings can be seen as assets, such as other financial assets, but they have some features that need closer examination.

The change in dwelling prices or rents occurs when the supply of dwellings increases or decreases with relation to the demand for dwellings in a specific submarket. Normally, the main drivers of the dwelling price change in a submarket are the number of available dwellings in the stock and changes in characteristics of the population in the submarket. Furthermore, the population’s fondness of a submarket may change due to various reasons which can affect the dwelling prices. Dwelling markets have their unique characteristics in terms of supply and demand since dwelling markets are markets for shelter and the dwellings itself are immobile. (Grigsby 1963). Immobility means that location matters to the prices of dwellings. Kiel and Zabel (2004) explain the three most affecting features to dwelling prices: location, location, and location. They elaborate that people care about their street or neighborhood, which is the first location. Great views, seaside location or good maintenance of the street, for instance, raise prices. The second location is a wider area where people live, for instance, a town. Facilities of the town, such as good health care services or quality of schools, have importance. The third location is the metropolitan area and its characteristics, e.g., temperature and cultural attractions.

Other typical units of trade are consumed and discarded when not anymore needed, e.g., food or phones, but when a family's need for a dwelling change, they have to find a new dwelling for living, which at the same time might mean moving from one area, or a submarket, to another. On the one hand, changes in families' aspirations can cause migration pressure and demand for dwellings to one submarket, and on the other hand, out-migration, and lack of demand for dwellings to some other alternative submarket, hence affecting contrary to the dwelling prices of the submarkets. Submarkets are interdependent, though, it is difficult draw a line where one submarket end and another starts. Facilities, e.g., a public transportation, stores or recreation areas have influence on a submarket's attractiveness. Submarkets with better facilities are more attractive, thus, submarkets are in constant competition. However, submarkets' attraction is relative and as discussed earlier, aspirations of population and other aspects might change and a better submarket in a specific time might be not so attractive in relation to another submarket at another time (Grigsby 1963). For instance, Covid-19 pandemic affected working culture and remote work became more common in relatively short period of time. Remote work and people's increased time spent in home caused a change in population's need for larger homes. In Finland in 2020, there was an 8% increase in moves from apartments to single-family houses in another municipality, and the trend continued in 2021 with a 6% increase (Huomo & Kannisto 2022).

Social cohesion is an important factor that improves residents' well-being, keeps societies and political environment stable and enhances cooperation between entities. High social cohesion may have a positive impact on the economy of society and vice versa (Dai & Sheng 2021). Dai and Sheng (2021) show in their study that an increase of uncertainty in the economy had a higher negative impact on real returns in the U.S.

dwelling markets in states where social cohesion was low. Therefore, one may argue that it is important to take different characteristics of an area into account when predicting prices of dwellings.

Ryan-Collins et al. (2017) explain a low-supply equilibrium in the UK, which is a negative feedback loop between homeownership and house prices. The number of homeowners has increased, and people's wealth is increasingly tied to the dwelling market. Housing prices would decrease if dwelling builders could supply more. Politicians need to balance between dwelling price affordability and sufficient supply, on the one hand, and on the other hand, they have an interest to maintain the homeowners' wealth. Governments have encouraged individuals to increase their personal wealth by investing in different assets such as dwellings to meet the costs of an ageing population. Multiple societies over the globe have moved towards asset-based welfare. However, the prices have risen so that more and more middle earning and poorer households cannot afford to own a dwelling and are forced to rent a home. Rents have also been at high levels, hence making it difficult to save for own home while paying rent of current home, therefore, situations shove some rent payers to the rent trap. In particular, this has been a problem among younger generations (Ryan-Collins et al. 2017). Ryan-Collins et al. (2017) argue that sooner or later problems will cause either a crash in dwelling prices which would reset the market or emergent frustration to larger rents would cause a political crisis with unclear consequences.

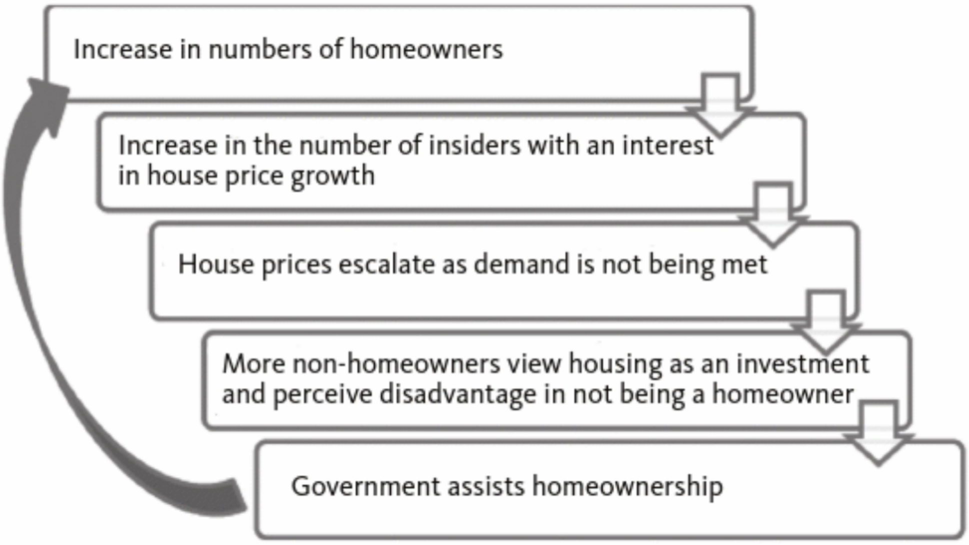
Increase in numbers of homeowners
Increase in the number of insiders with an interest in house price growth
House prices escalate as demand is not being met
More non-homeowners view housing as an investment and perceive disadvantage in not being a homeowner
Government assists homeownership

Figure 1 - Low-supply equilibrium (Ryan-Collins et al 2017)

In recent years, the mortgage credit market has increased due to cheap money available and relatively easy lending policies. Dwellings or properties are not only considered as homes but also speculative financial assets, which prices have grown steadily in the past decades. One of the reasons for dwelling price rise was explained already, but a second is lending against real estates which creates iterative credit supply, credit demand and rise of asset prices. This is called the positive house price-credit feedback cycle. The cycle operates in the following manner:

1. Mortgage lending surpasses the number of new dwellings and commercial buildings being constructed, hence resulting in an increase in property values.
2. Prices rise which forces households and firms to take larger loans, therefore, boosting banks' profits and capital.
3. With higher profits and capital, banks can grant more loans, which will increase property values (Ryan-Collins et al. 2017).

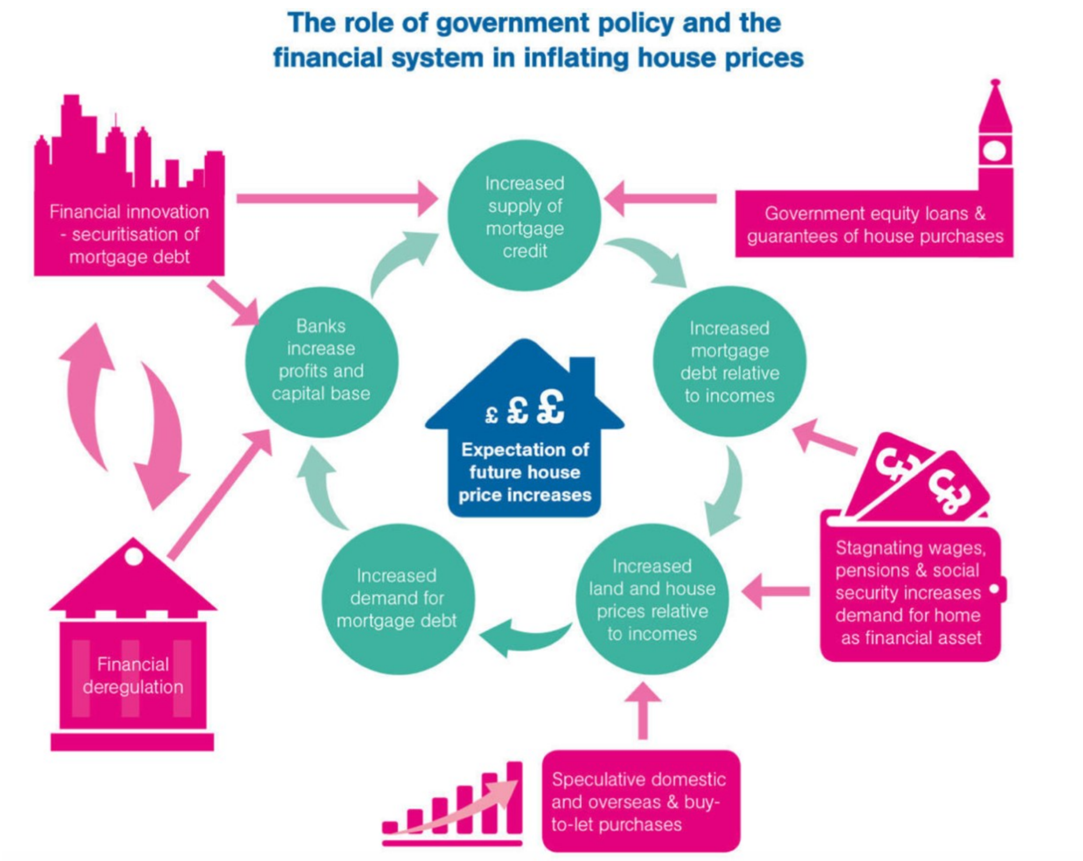
The role of government policy and the financial system in inflating house prices
Financial innovation - securitisation of mortgage debt
Increased supply of mortgage credit
Government equity loans & guarantees of house purchases
Banks increase profits and capital base
Increased mortgage debt relative to incomes
£ £ £
Expectation of future house price increases
Increased demand for mortgage debt
Increased land and house prices relative to incomes
Stagnating wages, pensions & social security increases demand for home as financial asset
Financial deregulation
Speculative domestic and overseas & buy-to-let purchases

Figure 2 - The house price-credit feedback cycle (Martin & Ryan-Collins 2016)

Increase of property prices can decrease commercial lending while banks concentrate on mortgage lending. This phenomenon can have a negative effect on investments of firms. The housing price-credit cycle can go on in kind circumstances but shock in the economy and/or tightening money policies can cause problems to indebted households, eventually, leading to falling prices or even a financial crisis (Ryan-Collins et al. 2017). The Euribor 12-month interest rate, the most commonly used rate in Finnish mortgages, has risen to 3.5% in early 2023. The change in the interest rate was rapid since the rate rose in a year prior to 6-year period when the rate was below zero. At the same time, the economy has faced supply shortages due to various reasons and in 2022, Russia invaded Ukraine causing devastating suffering to Ukrainians and also uncertainties to the economy. The future impact of the changed economic environment on dwelling prices, particularly in Europe, remains uncertain.

2.1 Finnish Dwelling Market

The Finnish dwelling sales market is dependent on overall market situation and institutional changes. During the last decades the Finnish dwelling market, as well as other markets, has faced ups and downs. For instance, from 1974 to the end of 1970's the real dwelling prices decreased due to the oil crisis, even though the nominal prices increased. In 1980's, Finnish regulators loosened the financial regulation, which made mortgage available for greater public and rapidly increased dwelling prices during 1987-1989. The rapid increase caused a boom in dwelling prices which eventually turned to decline mainly because of earlier rapid increase and deep recession in the early 1990's (Oikarinen 2007).

Finland is a sparsely inhabited country, with only 18 people living per a square kilometer. The population density is the lowest among the EU countries and most of the area of Finland is categorized in the rural category if the EU classification standards are used. Recent consolidation of municipalities has created vast municipalities which consists of urban and rural areas, therefore, the classification based on municipalities boundaries is not able to detect spatial differences of areas (Saastamoinen et al. 2022).

In the Finnish system, owning a property means that one owns a piece of land and possible buildings on the land. A common way to own a single-family home is to own a whole property. Though, there is different ways of owning a dwelling. A piece of land can be leased, typically from a municipality. In addition, a housing company can own a piece of land and individuals owning shares of that housing company. The shares entitle the owner to control a dwelling within the housing company. A block of apartments is a typical building where a housing company owns a piece of land and shareholders of the housing company are living in the company's dwellings. Taxation of transactions varies depending on which type of a dwelling entity is in question. Transaction of shares of a housing company causes a 2% transfer tax while transaction of a whole property causes a 4% transfer tax.

The quality of Finnish dwellings is relatively high, and the dwelling stock is relatively young since over 60% of the buildings were built after 1970. The mean size of the

Finnish dwellings is 80 square meters, and one person has averagely 40 square meters of a living space. Approximately two thirds of the Finnish population live in condominium dwellings while third lives in rental dwellings (Hannonen 2014).

The Finnish dwelling prices have been highly volatile past decades and the market has seen deep drops in recessions. One threat and a cause of uncertainty is the mortgage market, since 90% of the Finnish mortgages are tied to fluctuating market rates, such as the Euribor 12-months reference rate. The dwelling market itself is polarized. The Helsinki metropolitan and other growth centers lack of sufficient number of dwellings while other areas have high amounts of empty dwellings, and this phenomenon is going further. Polarization into good and bad subareas within cities is also seen in the Finnish cities. Some of the subareas have larger concentrations of unemployed people or immigrants, and some of the areas attract more wealthier people. The structures of the Finnish cities are scattered, and city residents tend to use cars (Hannonen 2014).

The urbanization megatrend is seen in Finland, too, and population in the large cities is constantly growing. In addition, immigration increases a flow to the Finnish major cities (Hannonen 2014). The COVID-19 pandemic had influence on the Finnish dwelling market and the results of the pandemic are still unclear. However, the urbanization seems to continue, but residents might make different choices between and within cities. Appeal of middle-sized cities and neighboring municipalities of large cities seems to be stronger. Especially, larger, and cheaper dwellings in those cities or municipalities have been appealing. However, as said earlier, the final outcome of the post-pandemic dwelling preferences is still unclear, and these trends might turn back (KTI 2023).

Legislation of Finnish rental agreements is one of the most liberal in the world. There are some regulations in the residential market, e.g., minimum time prior to a tenant must leave a dwelling when a landlord concludes the rental contract. Neither minimum nor maximum tenancy agreement terms are applied. An indexation is voluntary and contract renewal requires new agreements. The system is designed to give a freedom to contract parties to agree the conditions (KTI 2022). The rental market affects the total dwelling market, but this thesis will focus on the dwelling sales market.

Finnish dwelling market slowed down in 2022 after a peak year due to uncertain market conditions (Federation of Real Estate Agency 2023).

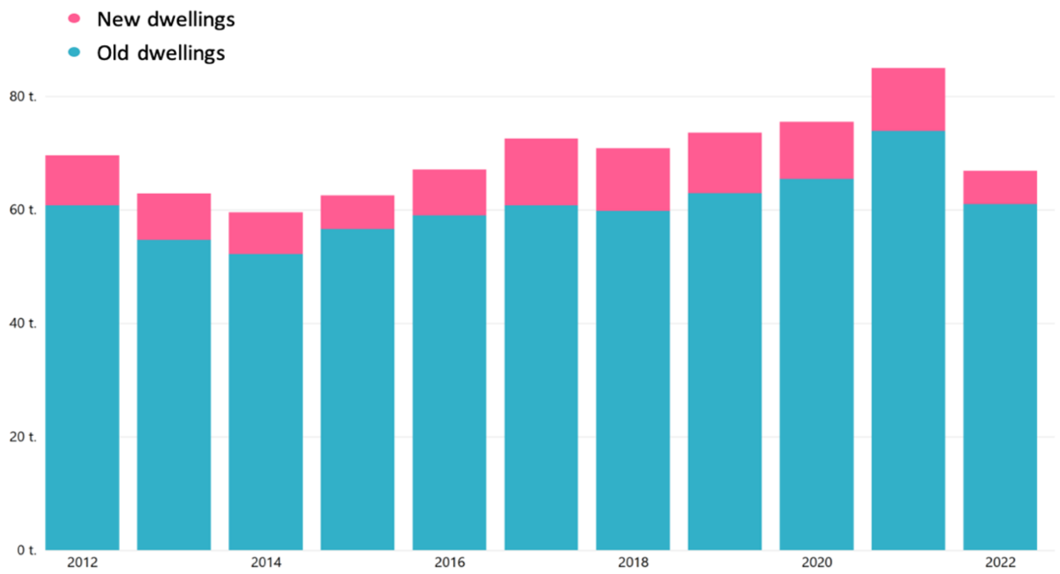

Figure 3 - Dwelling transactions pcs in Finland 2012-2022 (FREA 2023)

Value of the Finnish dwelling market dropped as well in 2022, after a peak year, and was 13.8 billion euros (FREA 2023).

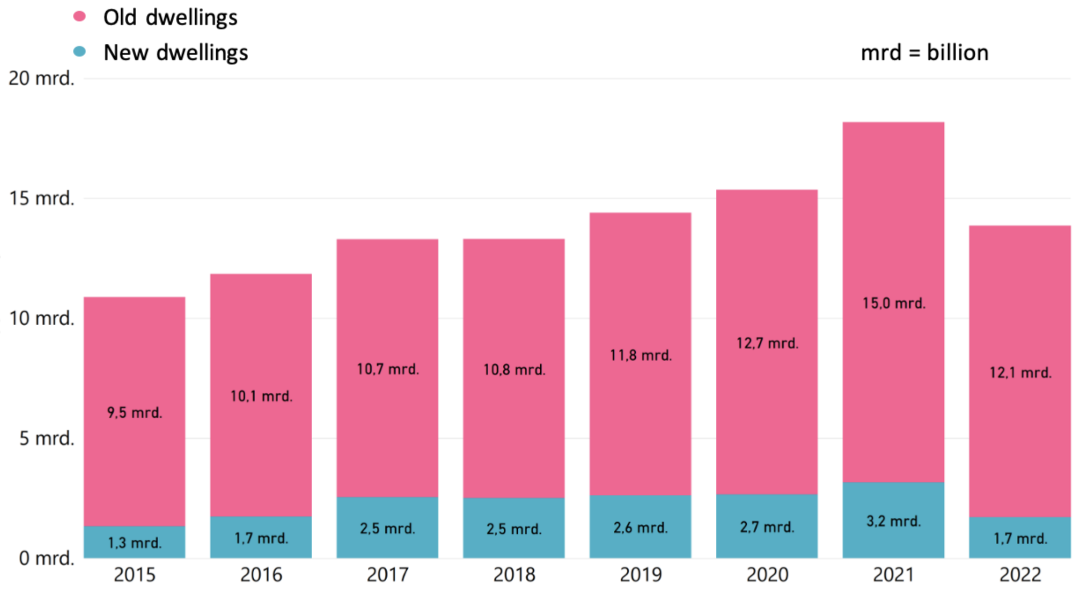

Figure 4 - Value of the Finnish dwelling market in 2015 – 2022 (FREA 2023)

The largest 10 city regions inhabit 62% of the population of Finland. The largest area by population is the Helsinki metropolitan area, which consists of cities of Helsinki, Espoo, Kauniainen and Vantaa, having 1.2 million residents. The second largest is the Tampere city region with circa 400.000 residents. The third largest population concentration is the Turku city region, which has around 340.000 residents. These Finnish regions are called “the growth triangle”. More than a half of the Finnish population lives in these regions and over a half of the jobs and gross domestic product comes from these areas (KTI 2023).

Espoo has a university, and it is the second largest city of Finland. It has a metro and railway connections and in near future a tram connection to Helsinki. Tampere is the largest inland city of Nordic countries, and it has two universities. It has a tramway and in recent years it has attracted lots of new businesses. Oulu is the largest city in Northern Finland and has its own university, too. Oulu has high influence on Northern Finland which mainly means a half of the whole country. Turku has Finnish and Swedish speaking universities, and it has good competence on sea cluster and biotechnology businesses. Jyväskylä, another university-city, is concentrated on traditional industries, e.g., wood and construction materials industries, and in addition ICT and healthcare industries. Kuopio has its own university and its business strategy concentrate on the food, health, bio, and environment industries. Unlike the other large cities in Finland, Lahti is the only one without university. Lahti has traditional industries, mainly on woodwork, furniture, and plastic fields (KTI 2023).

2.2 Dwelling pricing

Analysing real estate markets is difficult since there are two cognate markets in a market. The market for real estate as a space is determined by the needs of tenants which affects rent prices, while at the same time investors can do different transactions on real estates, when it is question of a real estate as an asset. When dwellings are owned or purchased by householders, there is no two markets anymore since purchasing an asset and the use of space are one single decision. The need of a real estate space rises dwelling prices if construction levels stay stable. When prices rise, construction is more profitable which encourage constructors build more dwellings, eventually satisfying the demand and prices fall back closer to construction costs (DiPasquale & Wheaton 1992).

DiPasquale and Wheaton (1992) produced a four-quadrant model that explains the real estate markets. Figure 5 explains the model. The model consists of four quadrants as the name suggests. The right-side quadrants present the property market for the use of a space. The left-side quadrants present the asset market for the ownership of a real estate. The North-East quadrant determines the rents in the short-term. The North-West quadrant determines a price, P, for real estate assets by taking the rent level R from the North-East quadrant and dividing R by the capitalization rate, which is the ratio of the rent to the price, in other words the interest rate that investors want for holding real estate assets. The South-West quadrant defines the portion of the construction of new real estate assets. The construction costs are replacement costs of real estates. The cost of construction is increasing if building activity increases. Different supply distractions make the supply difficult and moves the construction ray more horizontal. The price is given from the North-West quadrant and the construction line determines a level of the vertical axis. The level is the point where the replacement costs equal the asset prices, which means that the new construction levels that are lower than the point on the axis are profitable while the higher levels are unprofitable. The last quadrant, the South-East quadrant, is used to convert the annual flow of new constructions into a long-term stock of a real estate space. The depreciation line defines the level of the stock that needs an annual level of construction for replacement on that specific value on the change in stock axis. On that specific level on the stock axis the stock of space will stay stable. In short, the model defines rents based on the current level of property stock. The rents are converted into property prices. The prices define the needed construction which eventually defines the new level of stock. If the starting and the ending stock are in an equilibrium, the whole market is in a long-term equilibrium. If the ending stock is lower than the starting stock, then the rents, the prices and the construction should all rise to set the market in the equilibrium and vice versa. The model applies in the owner-occupied dwellings as well, nevertheless, the households' income level determines the North-East quadrant's 'Rent' and interest rates are used to determine the price (DiPasquale & Wheaton 1992).

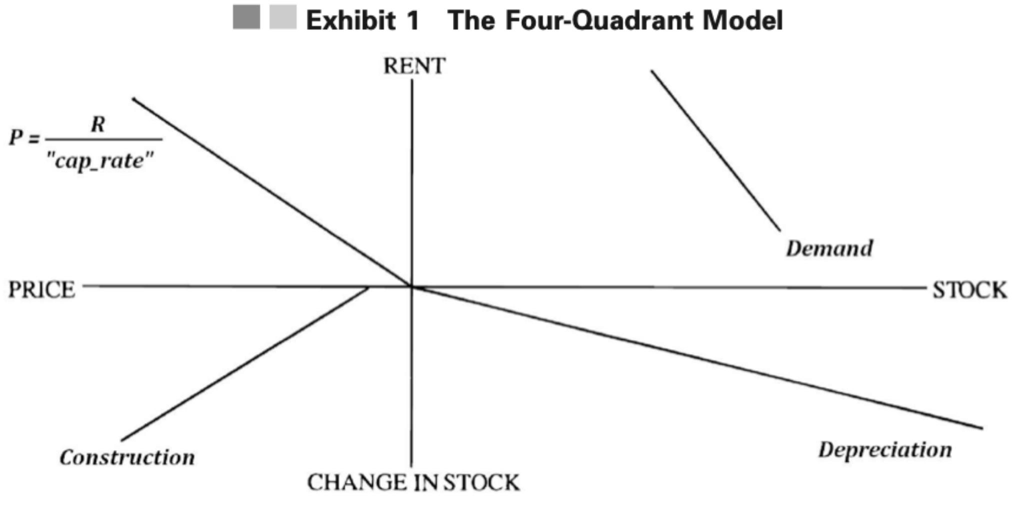

Figure 5 – The Four-Quadrant Model (Gaetano 2015)

Dwelling market cycles have been different in different cities. Some dwelling markets are more volatile than the others. This may be because of the initial distribution of debt levels of the cities. The cities where most of the residents' loan-to-value (LTV) ratios are high, tend to be more vulnerable to dwelling price crashes than the cities where the

LTV ratios are more diversified. Cities usually gather residents with high LTV ratios if the trading volumes are high during the rising price market period. Residents with high LTV ratios are less likely selling own homes compared to residents with low LTV ratios. Therefore, dwelling market booms may make the local dwelling market fragile, which can lead to a crash when a relatively small economic shock occurs (Stein 1995).

2.2.1 The Disposition Effect

Behavior of humans affects asset pricing. One studied phenomenon is the disposition effect. The famous phrase in stock market trading guides is “Cut your losses and let your profits run!” However, individuals tend to have difficulties to follow this guide, whereas they are willing to hold assets that are losing since the purchase and quickly sell the stocks that have risen (Kaustia 2021). The disposition effect is usually seen, when a seller is less sophisticated investor, such as household or non-profit institutions, while sophisticated investors tend to give less importance on the prior purchasing prices (Grinblatt & Keloharju 2001). This phenomenon affects the dwelling market among the other asset class markets. Dwellings are important part of household’s wealth and the dwelling market have significant effects to other areas of economy. Further, the dwelling market is less effective compared to stock markets, therefore, the disposition effect may have significant influence on welfare of the society (Kaustia 2001).

Dwelling prices and trade volumes have a high positive correlation. Furthermore, dwelling prices and time on the market have negative correlation (Genesove & Mayer 2001). Taking this into consideration, Einiö, Kaustia and Puttonen (2008) found that individuals tend to avoid selling dwellings, if they would lose money, especially in cases where they are selling low-priced dwellings and in cases when the seller has received social financing, e.g., state guaranteed loans. Social financing receivers usually have a smaller wealth, therefore mortgage down payments might cause constraints. However, that is not the only explanatory factor since the same phenomenon was found in cases where the mortgage constraint was probably not the issue, such as when dwellings were expensive or bought for investment purposes (Einiö et al. 2008). Genesove and Mayer (2001) found that sellers expecting losses tend to set higher asking prices compared to other sellers. That applied to investment dwellings, too, but with a smaller margin. Kaustia (2001) argues that the disposition effect is probably reason for much of the correlation between prices and trade volumes. In downturn, the decisions may be suboptimal for the dwelling market, which could lead to a weaker liquidity in the market and hamper the mobility of labor. Furthermore, the disposition effect increases individuals' capital taxes and reduces returns (Kaustia 2001).

Two things recommend a hypothesis that dwelling markets are inefficient. The seller characteristics, e.g., whether (s)he is loss averse or have equity constraints, have influence on dwellings' transaction prices. In addition, transaction volumes of dwellings decrease when prices decrease, which is not in line with perfect asset models (Genesove & Mayer 2001). One reason for declining transaction volumes is so called fishing strategy. Families that could move to other home but are not forced to sell the current one, can list the current home at a higher price than the market price in a declining market and just wait if they are lucky and someone buys the home. If no-one buys, they can keep the current home and forget the new potentially more suitable one (Stein 1995). Reasons why people are reluctant to realize losses may vary, but one could be that they comprehend unrealized losses as only paper losses and do not admit the fact the investment has been poor (Kaustia 2001).

2.2.2 Features of the Dwelling and Spatial data

Dwelling characteristics affect sale prices. For instance, Jim and Chen (2006) found that the high floor level in a multistorey building has a positive influence on the price of a dwelling. Larger dwellings are usually more expensive, while a specific heating system might have either a positive or a negative impact. Typical features used in hedonic pricing models are, for instance, number of rooms, floor area, building type, heating system, age, constructing materials, and other structural features, e.g., sauna, basement, and garage (Malpezzi 2003).

Energy performance is a hot topic at the time. In 2022, Europe faced energy crisis caused by a lack of natural gas due to the geopolitical situation. The crisis increased households' energy costs dramatically during 2022, but a mild winter threw the worst anxiety away. However, the uncertain geopolitical situation is likely to continue, therefore, the energy performance has been and will be an important feature of a dwelling still in the future. Fuerst & Warren-Myers (2018) found that energy-efficiency ratings and other sustainability-related characteristics have an impact on the pricing of dwellings. Buyers are willing to pay more of energy efficient dwellings (Brounen & Kok 2011, Koengkan & Fuinhas 2022).

The traditional hedonic pricing models do not take into account spatial heterogeneity of areas. After 1990's the effects of spatial heterogeneity, i.e., neighborhood differences, have been in the spotlight and have been noticed that ignorance of spatial heterogeneity might cause biased estimations for dwelling prices (Wu 2019). Dwellings' individual characteristics are important, but it is broadly accepted that an area's social-economic conditions and location aspects matter. An area's income level, median age, education level and distance to a city center, for instance. Socioeconomic circumstances of a neighborhood have connection with residents' affluence, therefore affecting the prices of dwellings in that submarket. Typically, a location has lots of value due to its relation to accessibility, commuter transport and living environment (Wu 2019).

Some of the spatial features are not so clear. Osland & Pryce (2012) argue that distance to employment and dwelling prices have nonmonotonic relationship. When a dwelling's distance increases from an employment concentration, i.e., office buildings, the price of the dwelling initially increases as well, but when the distance is far enough the price starts to decline.

Fuerst and Warren-Myers (2018) claim that socioeconomically disadvantaged districts suffer from excessively higher levels of energy-efficiency rating non-disclosure which can lead to the green lemons problem. The green lemons problem refers to the unsymmetrical information between buyers and sellers that might affect the pricing of a dwelling (Fuerst & Warren-Myers 2018). In addition, research has shown that in Finland, the quality of schools has an impact on dwelling prices (Pakarinen 2018). Leech and Campos (2001) found that the school districts of popular schools have a positive effect to the dwelling prices in the UK. These factors have crucial impact on how dwelling prices are formed and the overall performance of the dwelling markets (Cellmer et al. 2020), therefore them should be included in dwelling price prediction models.

3 MACHINE LEARNING

Machine learning has rapidly emerged as a critical tool for analyzing large datasets and predicting future trends. In recent years, it has found increasing use in the field of real estate, where it is being used to predict dwelling prices, analyze market trends, and identify investment opportunities. The chapter offers a deeper understanding of machine learning and its various methods, e.g., supervised and unsupervised learning. The fundamental principles of machine learning will be explored and how it can be used to predict dwelling prices. The chapter will present an overview of the different types of machine learning algorithms and the key steps involved in building a machine learning model. The information discussed in the chapter is crucial for the thesis, it is valuable for the reader in order to understand the process and the results of this thesis. In addition, the chapter guides the author with the machine learning modelling process.

Today's world is full of data. Not only large companies are generating data, but we all are. Every time we buy something, read some news on a website, or send a cute cat photo on the social media, for instance, we are producing data. Furthermore, we are not only producers of data, but we are consumers of data. We want that our phone applications or just companies understand what we want and make our lives effortless. Individuals' behaviour and other subjects in the world often contain patterns, and these patterns can be learned from sufficient amount of data. When we know a pattern, we can create an algorithm which can turn an input to an output. Some of the tasks might be so complicated that we know the input and the output, but we are not capable of making sense of the algorithm, in other words, how to turn the input to the output. We can give that task to a computer, which can learn from data and create the needed algorithm. The understanding of the created patterns helps us understand the process and enables us to make predictions (Alpaydin & Bach, 2019)

Simply, machine learning means an assortment of various computer algorithms and techniques, which are used to produce systems that are capable of learning from data. The systems can solve difficult problems from a specific domain, such as detect deceitful credit card transactions without understanding of the domain knowledge itself, whereas relying on mathematics and statistics (Wei-Meng, 2019, pp 3).

Machine learning can solve different kinds of problems, e.g., classification, regression, and clustering problems. An answer to a classification problem illuminates if something is A or B, e.g., based on data, a model can tell that the observed object should be an apple, not an orange. Regression models answer to questions how much or how many, e.g., predict a sale price when certain factors of an object are known. Clustering models organize data to natural groups and help understand them, e.g., which groceries are commonly bought together (Wei-Meng, 2019, pp 5).

Dwelling price prediction is seen as a regression problem since we try to predict how much a dwelling should cost based on its features. That is the reason why we focus, in this paper, on models that are capable of solving regression problems. Moreover, the dwelling price prediction has non-linearity features since some of the features of dwellings are non-linear. A building year of a dwelling can be seen as an example. Old value dwellings, which are built in the early 1900s and newly constructed dwellings are typically more expensive than dwellings which are built 1960 – 1990, for instance (Kalliola et al. 2021).

As discussed earlier, machine learning can be divided into different learning types, which are discussed next.

3.1 Supervised Machine Learning

Supervised machine learning is in question when a dataset with labelled features is used, e.g., a dataset contains information regarding dwellings features and prices (Wei-Meng, 2019, pp 6). As the name suggests, the supervised machine learning system has been guided by a supervisor into use of labels and associate them with training examples (Cunningham et al., 2008, pp 21). A model taught by supervised learning is capable of predicting an output when it receives previously unknown input information (Alpaydin & Bach, 2019, pp 21), e.g., receives information on an apartment and then predicts its price.

Training data are rarely, if ever, comprehensive to generate a unique model which predicts perfectly all the possible situations. That means learning is ill-posed and requires additional assumptions leading to inductive bias. That is not evitable since all our models are inductively biased on some way and the task to do when creating a good model is to choose the right bias and find the best generalization. When creating a model, we try to generalize it as well as possible in order to make it work well with new data.

Underfitting means that the model is not capable to find trends of the training data, therefore, the model cannot understand the relationships between input and output data and performs poorly. For that reason, the model should be complex enough to find the needed trends, but not too complex, which can lead to overfitting problem. Overfitted models cannot make accurate predictions either since they typically have learned the noise and imprecise data inputs of a training dataset causing a poor generalization. The following points should be in mind when creating a model. Typically, the size of a dataset matters. For example, the generalization error decreases if we have more training data. Moreover, the generalization error decreases when the complexity of a model increases, albeit only to a certain point when the complexity of a model starts to increase the generalization error (Alpaydin 2014). Supervised machine learning techniques are usually used for dwelling price prediction tasks and this study will discuss them more in following chapters.

3.2 Unsupervised Machine Learning

Another type of machine learning is called unsupervised learning. Unlike supervised machine learning, where input and output features were labeled and the goal was to find the input and output pairs, unsupervised learning aims to find meaningful patterns, associations, relationships, or clusters without non-pre-labeled data and generate some knowledge from them. The problem is not so clear since there are no axiomatic patterns to look for or clear error metrics to use (Murphy 2012; Subasi 2020).

The training data go same kind of pre-processing steps than in supervised machine learning, but the aim differ from supervised learning, i.e., feature selection tries to find most informative features for clustering tasks and normalization is done to make the data centered. In the prediction step, a validation or a test set is used to evaluate the model (Subasi 2020). Examples of unsupervised learning problems are customer segmentation and market basket analysis.

Unsupervised learning can be used in the field of dwelling markets. Ntantamis (2010) used an unsupervised machine learning method to identify different submarkets based on their spatial characteristics. Kim and Irakoze (2023) utilized unsupervised learning to extract dwelling transactions with green certificates from other transactions. Examples of unsupervised learning techniques are k-means, affinity propagation, mean shift, gaussian mixture models and dimensionality reduction. K-means is a popular clustering algorithm which simply separate data to different clusters by minimizing the squared distance between the cluster mean and single observations in the cluster. The user needs to set up the number of clusters needed (Subasi 2020). It can be used, for instance, to group similar dwellings based on their features, such as price, location or building type.

3.3 Reinforcement Machine Learning

The third type of machine learning approach is called reinforcement learning, though, it is also considered as a paradigm in supervised learning (Subasi 2020). However, this type of learning relies on reward and punishment signals which affect following actions or behavior of the model (Murphy 2012). Individual actions or behaviors are not important but the policy, which makes the model correct its actions or behavior to reach its goal, matters. The policy is achieved through past experiences of trial and error (Alpaydin 2014).

Reinforcement learning algorithms are useful when the goal is clear, and a machine learning model should learn to interact with its surroundings in order to gain the goal. Reinforcement learning can be used, for instance, on dwelling’s energy or lightning management. The algorithm would receive either positive or negative signals – rewards or punishments – based on the dwelling's energy consumption and living comfort, thereafter, would adapt heating or lightning accordingly to those signals in different conditions.

Diyan et al. (2020) employ reinforcement learning to enhance energy management in a smart home. Reinforcement learning is used in a study that finds mortgage borrowers who have paid partial prepayments, have higher probability to pay prepayments in the future compared to borrowers that do not have such experience (Deng et al. 2021). Park et al (2019) utilized reinforcement learning to improve lightning management in terms of occupant comfort and energy efficiency.

3.4 Pre-processing

Machine learning models are as good as their input data. Therefore, the data pre-processing is an important task to do prior to the model training phase. The pre-processing phase is usually time-consuming, but the results of the models are more accurate, when this step is done well. This chapter will discuss on data pre-processing steps in order to make data useful for machine learning models.

3.4.1 Data Integration

Typically, raw data are gathered from different sources, and it is stored in multiple different datasets. To make data usable and valuable for machine learning model, the different datasets should be integrated or merged. Integrating means that datasets of similar features are combined while merging means that datasets with dissimilar features are used to supplement each other (Subasi 2020).

3.4.2 Missing data

Data are irreplaceable asset but only if the data are accurate and conclusive (Alruhaymi & Kim 2021). Most of the real-life datasets are incomplete, in other words, they have missing data for various reasons (Brownlee 2020). Dealing with incomplete data, is a significant challenge, when building a proper machine learning model. Missing data means values that should exist but are not available. Many statistical and machine learning techniques are not able to handle missing values properly, which can cause uncertainty and bias in the analysis. Ultimately, that can lead to inaccurate conclusions and decreased performance (Jadhav et al. 2019, Brownlee 2020).

Missing values can be handled two ways. The missing data can be ignored; thus, the features or observations with missing values can be totally removed from the dataset. This approach is problematic since it decreases the size of a dataset (Jadhav et al. 2019). The second way is imputation of missing values, in other words, to fill the missing values with some probable value, which will eventually reduce bias because missing information decreases and valuable information increases (Jadhav et al. 2019). For instance, the filling can be done with a mean value of all the known values of a feature in question. Must notice, that if a feature has lots of missing values, it might be impossible to fill the missing values with plausible ones, therefore, removing a feature might be the best solution in certain cases.

The imputation of missing values can be done multiple ways. The traditional ways are explained next. Imputation with constant means that the missing value is filled with a constant, for instance, with a value ‘0’ or ‘good’. Mean, median or mode imputation refers to a solution where the missing value is filled with a mean, median or mode of sample values. Mean imputation can change the shape of feature’s distribution (Jadhav et al. 2019) and cause bias due to decreased data points’ variance (Alruhaymi & Kim 2021). However, the effect is slight if max 10% of the data are missing and the correlations between features are not too high. The imputation of missing values can also be done by using imputation with distributions, regression imputation or KNN imputation techniques, which will not affect the shape of the distribution. In imputation with distributions, the missing values are changed by a randomly chosen value from a known distribution. Regression imputation means a technique where other variables are used to predict the missing value. KNN imputation is a method where values are filled by copying values from a same kind of records in the same dataset. However, this is a time-consuming way (Jadhav et al. 2019).

Missing values can be categorized to three categories, which are Missing at Random (MAR), Missing Completely at Random, (MCAR) and Missing Not at Random (MNAR). Classification is important, since it affects ways, we can deal with missing values. If data are considered MCAR, it means that there is no relationship between the missing values and observed values, thus, it can be assumed that no bias occurs in the data remaining. This type of missing values is the easiest to deal with. The values can be ignored without fear of a biased dataset or traditional imputation methods can be used. Same applies to MAR. MNAR data are due to factors of missing data, meaning that the missing data are not a random occurrence, therefore, it is the most problematic type of missing values. Removing observations containing MNAR data may lead to a biased dataset (Alruhaymi & Kim 2021). However, the categorization is not always clear and sometimes the character of missing data could be mixed. In addition, the detection of MNAR can be difficult since it is not always apparent that the missing values have relationship with collected data.

3.4.3 Outliers

Detection of outliers is an important task when building a machine learning model. Outliers are data points that are significantly different from the other observations in a specific dataset. Patterns that do not follow a normal behavior (Singh & Upadhyaya 2012). The definition might vary depending on the data structure. Generally approved definition is Hawkins' (1980) definition "*an outlier is an observation that deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism"* (Yang 2020).

Unusual or deviant patterns or observations in a dataset are outliers (Singh & Upadhyaya 2012) but there is no explicit answer which values are outliers since every dataset has its own characteristics (Brownlee 2020). Outlier detection process tries to identify these outliers which do not follow the expected patterns or observations in the dataset. Normal observations form an own region or regions and the observations out of that region are considered outliers (Singh & Upadhyaya 2012). Outliers typically have negative effects on statistical analysis, they can have biased influence on important estimates, therefore, lead to imprecise conclusions. The variability of the data may raise due to outliers and hinder detection of meaningful patterns. In addition, the normality of the data might suffer if outliers are not randomly distributed (Osborne & Overbay 2004).

Many reasons can cause outliers to a dataset. Errors can occur in data collection, recording and entry due to human errors (Osborne & Overbay 2004, James et al. 2017). For instance, a realtor can accidentally input data of a dwelling incorrectly to a data system which would have an impact on the results of this thesis. Of course, a realtor could do this purposefully, too. Motivated mislead reporting might occur when a feature is socially desirable. Sampling errors, standardization failures and defective distributional assumptions can lead to problems (Osborne & Overbay 2004).

Detection of outliers might feel simple, but it can be challenging for various reasons. The defining of normal pattern or behavior is cumbersome. For instance, detecting every normal pattern or behavior as normal. The boundaries between normal and abnormal behavior might vary and normal behavior usually evolves, therefore, abnormal behavior today can be normal tomorrow. Application domains are different and considered outlier values in one domain are not universal (Singh & Upadhyaya 2012).

Simple rules can be used to detect outliers and decide what to do with them. For instance, if observation is three or more standard deviations from the mean, it can be considered as an outlier, though, this can be problematic if a distribution of a small dataset is very skewed (Osborne & Overbay 2004).

A dataset can contain various types of data. The nature of data determines usable outlier detection techniques. Statistical techniques can be used continuous and categorical data while some attributes require nearest neighbor or distance measure techniques (Singh & Upadhyaya 2012).

The simplest outlier is a point outlier which refers to a single data observation which varies from the rest of the data and is therefore considered as an outlier. For instance, in the dwelling market, there can be a house which is significantly larger, let us say 1,000 square meters, while normal dwellings are somewhere around 30-200 square meters. The 1,000 sqm dwelling is considered as a point outlier (Singh & Upadhyaya 2012).

Contextual outliers are data observations which are abnormal in a specific context, but not in other. Individual observation is defined by contextual and behavioral attributes, where contextual attributes are location in spatial data or time in time series data, whereas behavioral attributes determine non-contextual attributes. For instance, EUR 1,000,000 (behavior) priced dwelling in the city of Savonlinna (context) can be considered as a contextual outlier, but the same priced dwelling in the city of Helsinki could be considered as a normal data observation. Collective outliers are data observations which are not itself outliers, but they are outliers as a collection (Singh & Upadhyaya 2012).

There is no exact answer what to do with outliers. If a dataset contains unlawfully added data, those should be removed. However, the detection of unlawfully added data can be extremely difficult or even impossible. Some researchers say that legitimate or unclear outliers should be removed, too, to achieve the most accurate estimates of population parameters. Some researchers disagree. One way to keep the extreme data points is to transform the data, for instance, transform the data to logarithm scale, which would decrease the skewness of a dataset (Osborne & Overbay 2004).

3.4.4 Feature Engineering

Some features have valuable information to a problem at hand, but not in the original form. That kind of features can be produced to new features which are more valuable. The process is called feature extraction (Subasi 2020). A new feature can be a modification of an original feature or multiple features. For instance, a dataset has a feature which specify the number of crimes in an area and a feature which specify the number of residents in an area. Those features might be useful for a certain problem but for another problem a crime rate ratio might be a better feature. The two features in hand can be used to produce a crime rate ratio.

Usually, machine learning algorithms work best with numbers. Categorical features with text might be problematic in that case and they should be converted into numerical features. One way to do that is to use OrdinalEncoder which converts text values into numerical values by giving values numerical labels. Machine learning models assume that numbers close to each other are similar and numbers far from each other dissimilar. Converting categorical values into numerical values with OrdinalEncoder can cause problems with certain variables if the distance of values does not represent their similarity. This problem can be handled with one-hot encoding, which creates new binary variables from individual categories within a categorical feature. The new independent variables are called dummy variables. If an observation belongs to a specific category, that specific dummy variable has a value 1, and the specific dummy variable has a value 0 if the observation does not belong to that category (Aurelien 2019). Example of a categorical variable is a city variable that contains information whether an observation belongs to a specific city. The variable can be converted into dummy variables which represent individual cities. If an observation belongs to a specific city, e.g., Tampere, the dummy variable Tampere has a value 1 but the other dummy city variables have 0.

Many machine learning algorithms assume that variables they encounter are nearly normally distributed. However, in real world cases, it is typical that some of the variables have a tail-heavy distribution, thus, training a model with a tail-heavy distribution may lead to biased predictions. This sort of variables can be transformed, for instance, by computing the variable’s logarithm, which usually transforms the variable to more kind of normally distributed (Aurelien 2019, Subasi 2020).

3.4.5 Feature Selection

Datasets might have information which is not needed in a problem at hand. Usually, that information should be removed and only needed data chosen into the final dataset. Features are selected based on their quality and predicting importance.

3.4.5.1 Collinearity & Multicollinearity

If two independent variables have high correlation with each other it is a question of collinearity. Collinear independent variables can cause problems to regression models since their individual influence on the target variable is tough to determine. Collinearity decreases accuracy of the estimates of the regression coefficients and likelihood of detecting non-zero coefficients. Independent variables with high correlation can be detected by using correlation matrix and searching independent variables with high absolute values (James et al 2017).

Collinearity is a problem in the dwelling price prediction since many of the features are correlated with each other, for instance, as the size of a dwelling increases, typically the number of rooms increases as well (Pakarinen 2018).

The correlation matrix does not reveal all the collinearity problems since collinearity can occur between multiple variables, even if two independent variables do not have collinearity between themselves. This situation is called multicollinearity and it needs its own measurements to detect. One way is to determine multicollinearity by calculating the variance inflation factor (VIF). Usually, the independent variables have some collinearity, however, if the value of VIF exceeds 5 or 10, it can be considered problematic (James et al 2017).

There are simple solutions to the collinearity problem. The problematic independent variable might be removed. Usually, the removal of an independent variable with high collinearity does not jeopardize the fit of regression due to the fact that the variable does not bring lots of additional value on top of other variables. The second solution is to create a new variable from two correlated variables (James et al 2017).

3.4.6 Feature Scaling

Features used in a machine learning model might cause biases to the model if the scales of the independent variables vary a lot. Therefore, features typically need scaling and normalizing measures (Subasi 2020). Data used to predict dwelling prices, typically, contain different variables with totally different scales. For instance, prices are shown in thousands or even millions while square meters of a dwelling are typically from tens to hundreds or different Boolean variables with values 0 or 1. Prediction models might see data with smaller values insignificant and focus on variables with higher values, which would be problematic. Because of that, the deployment of feature scaling actions to a dataset is important (Huang & Le, 2021 p39).

Data normalization, or in other words min-max scaling, means that the values of different variables with different scales are scaled such that they are ultimately confined within the numerical range of 0 to 1. The process is done such that a value of a feature is subtracted by the feature's minimum value and divided by the maximum value minus the minimum value. The Scikit-Learn has its own min-max scaler which has hyperparameter that allows a user to change the scale from 0-1 to something else if that is needed (Aurelien 2019).

Standardization is another way to change scales. Standardized values have always 0 as a mean value. In standardization process, the mean value of a feature is first subtracted from all data points of the feature. Then the results are divided by the feature's standard deviation. The standardization process does not result in any specific range unlike the min-max scaling. The results that are outside of three standard deviations from the mean can be considered as outliers. Therefore, the standardization is more resilient to outliers than the min-max scaling (Aurelien 2019).

3.5 The Bias-Variance Trade-Off

In supervised machine learning, two major objectives are constantly competing with each other. On the one hand, models are supposed to fit the training data as well as possible, in other words, find as low bias as possible, while on the other hand, the objective is to generate models that are generalized and which work well with new data, in other words, has low variance (James et al. 2017, Rogel-Salazar 2018). High bias models underfit the training data which means that they cannot find meaningful relationships between inputs and outputs and have high errors with both training and test data, whereas high variance models overfit the training data which means that the models remember the training data too well and are not generally meaningful, resulting poor performance with new data. This competition is called the bias-variance trade-off (Rogel-Salazar 2018).

As a rule of thumb, the more flexible methods have higher variance and lower bias. To achieve good test set performance results, a model should find a balance where the variance and the bias are low (James et al. 2017). Reducing independent variables can lower the variance and a decision to keep the strongest predictors – independent variables with high absolute correlation with the target value – can lower the bias. Effective solutions depend on the learning task. In general, the risk of high bias is increased when a dataset is small, and the variety of different situations is high (Ehrig & Schmidt 2021). Possible solutions in order to prevent overfitting and inappropriate complexity are different regularization methods, i.e., adding some constraints how the model learns. (Rogel-Salazar 2018).

3.6 eXplainable Artificial Intelligence

Machines learn effectively from the given information, develop their own patterns to solve fast high-complex tasks, and therefore help humans' life to develop. Sometimes machines develop their complex models almost independently, without a human touch, which has led to a situation where humans are lacking, at least in some cases, of understanding of these models. We use AI systems in sensitive fields, too, such as medicine and defence. Therefore, it is important to be capable of explain how AI systems do their decisions and how they perform in certain situations in order to ensure the decisions created by models and used by users, humans or machines, are well-grounded and rightful (Arrieta et al. 2020).

Linear Regression, Decision Tree, K-Nearest Neighbours, Rule-based Learning, General Additive, and Bayesian models are transparent by their design, which makes them automatically understandable. However, sometimes also these models, depending on various reasons, can be difficult to understand, especially if the audience of a model is not familiar how it works. More complex models, which utilization degrees have risen recent years due their empirical success, e.g., Tree ensembles, Support Vector Machines, Multi-Layer Neural Network, Convolutional Neural Network, and Recurrent Neural Network models are highly complex by their nature and therefore difficult to understand (Arrieta et al. 2020). Usually, simpler models are understandable and more plausible (Alpaydin 2014), but they are lacking performance when compared to more complex models. Furthermore, enhancing the complex model’s understandability can reduce the model’s performance. Nevertheless, it is important to notice that more complex models are not automatically better performing ones, e.g., in cases when data are well structured, and features of sample data are representative (Arrieta et al. 2020).

Transparency of machine learning models is improvable; however, it usually means balancing with accuracy of the models. Post-hoc comprehensibility techniques can be used to open the functions of complex models and understand their reasonings behind the decisions. Post-hoc techniques can be algorithms which are designed to specific machine learning models or to all kind of machine learning models. Albeit these techniques are designed to complex models they can be applied to simpler ones too if that is needed. For instance, if the audience is not necessarily an expert of a field, eventually objectively assertive explanations are needed. Good explanation introduces why model made a specific decision instead of another option (Arrieta et al. 2020).

Model-agnostic techniques are designed to work with any machine learning models. Different techniques fall under the umbrella of this title. Explanation by simplification tries to make the original model simpler to understand by extracting rules, for example.

Feature relevance explanation tries to reveal an originally cumbersome model's features and their role in the prediction. Visual explanation techniques can be used, yet often with other techniques. Hybrid methods combine various techniques, i.e., transparent model and complex model can be paired in order to enhance understanding of the complex model, or knowledge of the complex model can be improved by utilizing knowledge of a transparent model (Arrieta et al. 2020).

This study aims to understand the features of developed models; therefore, only interpretable models are used in the empirical part. Different machine learning algorithms are discussed next.

3.7 Machine Learning Algorithms for Dwelling Price Prediction

Next, we will discuss different machine learning algorithms. The No Free Lunch Theorem suggests that there is not any single learning algorithm that is superior in some specific domain, hence different algorithms should be always tested and select the best performing one (Alpaydin 2014). Due to nature of this study, only interpretable models are discussed in this section.

3.7.1 Linear Regression

Linear regression is a simple way to conduct supervised learning. Linear Regression is generally used for forecasting and understanding quantitative relationships between independent variables and a dependent variable. It is used when the dependent variable is continuous and only in supervised learning problems. It is not suitable for categorical cases. Such as the name suggests, a linear regression model tries to find a linear relationship between independent variable(s) and the dependent variable. A simple linear regression contains only one independent variable whereas models that explain dependent variable by using multiple independent variables are called multiple linear regression models (Manasa et al. 2020). Even though, the linear regression is clearly a weaker approach than many other newer algorithms, it is still useful, widely used and a good starting point for a model development (James et al. 2017). The formula for a simple linear regression is

$$y = a_0 + a_1 * X$$

where a_0 represents y-intercept (the value of y when X is zero), a_1 is the coefficient of independent variable X. The simple linear regression is really an unfussy approach. Y could be a sale price of a dwelling, for instance, and X square meters of a dwelling. Linear regression model tries to determine estimates for the coefficients or in other words for the parameters (a_0 and a_1 together) from the training data. After the parameters are estimated, we can try to predict estimated sales price ŷ for the future observations. In real world, linear models typically miss some of the relationship, due to the fact that the relationships are not usually linear. There can also be some other variables, for instance, a condition of a dwelling which causes some variation. Therefore, we should add the error term ε to the model, which explains that random noise (James et al. 2017).

Simple linear regression is powerful if a dependent variable can be explained by using one independent variable and they have a linear relationship. However, in practice problems generally contain more than one independent variable, therefore the multiple linear regression is needed. The formula for multiple linear regression is

$$y_i = a_0 + a_i * X_i + a_{i_2} * X_{i_2} + \cdots + a_{i_n} * X_{i_n}$$

where n means total number of independent variables. A multiple linear regression model takes several independent variables into account, for instance, size, condition, building year, and heating system. Again, the coefficients are unclear, and the model needs to estimate them to produce a prediction ŷ (James et al. 2017). Important to note, the multiple linear models are not able to use the heating system variable straightaway, since it is a categorical variable. The variable can be coded to numerical, by using one-hot encoding, for example.

The linear regression models assume that the independent variables and the dependent variable have a straight-line – linear – relationship. If the relationships are not linear the prediction accuracy can be poor (James et al. 2017).

Zhang (2021a) uses multiple linear regression for the dwelling price prediction and concludes that multiple linear regression can be used for that purpose but with a limited accuracy. Liu (2022) achieves maximum of 7.6% dwelling price prediction error by using multiple linear regression in China. Anand et al. (2021) achieve 86% dwelling price prediction accuracy by using only four independent variables. Jiang & Qiu (2022) predict prices in 31 main cities in China and achieve 0.947 R-Squared by using two independent variables. Whieldon and Ashqar (2022) predict dwelling prices in Catonsville and multiple linear regression has 0.98 Adjusted R-Squared. Previous studies show that linear regression is useful tool for dwelling price prediction.

3.7.2 Decision Trees

Decision trees are tree-like hierarchical decision models used in supervised learning to either classification or regression problems. Decision trees are nonparametric methods which means that they do not have assumption about underlying distribution of the training data. Instead, the independent variable data space is divided into local regions, for instance, using the Euclidean norm, and then that local region is used for every individual observation (Alpaydin 2014).

Decision trees consist of decision nodes and terminal leaves. Every decision node executes a test function and gives an outcome to separate branches. When decision tree receives an input, the decision-making process starts from a root node. After the root node has executed its function and one of the two (though, a tree can have more branches if a multivariate decision tree is used) branches is selected, the function of the selected branch is executed next. This process goes on as long as a leaf node is reached. Every leaf node has a label which is the output value for the input. In case of the dwelling price prediction leaf nodes get a value that determines the price of an input dwelling (Alpaydin 2014).

In the tree's growing process, decision nodes split the training data into smaller groups until the two resulting groups are as homogeneous as possible regarding the response y. To put it another way, decision trees split the data and try to reach the smallest variance within the resulting groups. The estimate of the decision tree's regression function at each resulting leaf node is the mean value of the dependent variable which is constructed from the samples in that specific node. The mean value represents the prediction of the dependent variable and is used for all new observations that end to the node in question (Hastie & Tibshirani 2017). The structures of decision trees are not fixed. Trees grow, in other words add more branches and leaves, depending on the complexity of the problem in training data (Alpaydin 2014).

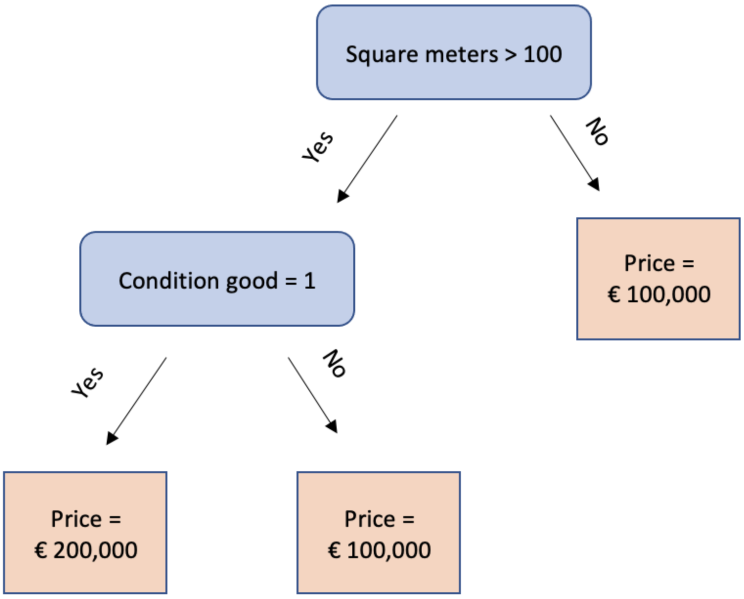

Figure 6 - Example of a simple decision tree

Decision trees tend to overfit if trained too thoroughly, which causes a poor test set performance. The tree learns the training data too precisely, becomes too complex, and is not able to generalize itself. One solution to solve the overfitting problem is tree pruning. It can be done so that a tree is allowed to grow only as long as the decrease in the residual sum of squares due to each split crosses some threshold value. However, this approach usually stops the tree growth too early, and the performance suffers too much. Since, a better way to do pruning is first let a tree grow large and prune it back by seeking a subtree. The best subtree can be chosen by conducting, for instance, cost complexity pruning (James et al 2017). The decision trees are popular since their interpretability is good, but they are not the most effective ones (James et al. 2017). Alpaydin (2014) suggests that a decision tree should be tested and take it as a benchmark before more complicated algorithms are used.

Decision trees have been used in dwelling price prediction projects. Louati et al. (2022) achieve better dwelling price prediction results with a decision tree than with the linear regression model. Kayakuş et al (2022) predict Turkish dwelling prices and the decision tree model outperforms the support vector machine and the artificial neural network models slightly. R-Squared is impressive 0.989. Tekin and Irem (2022) test different models and predict dwelling prices in Istanbul. The decision tree model achieves the mean absolute error of 25.42% and outperforms the linear regression and the polynomial regression models. Zhang (2021b) predicts dwelling prices in Boston and the decision tree model outperforms the linear regression and support vector regression models. Fan et al. (2006) predicted dwelling prices in Singapore and the decision tree model achieved 0.885 R-Squared. Peng et al. (2019) predicted dwelling prices with a decision tree model and that outperformed the multiple linear regression algorithm, yet it underperformed the more advanced model.

3.7.3 Random Forest and ExtraTrees

Typically, the models that combine multiple learners are more effective than individual machine learning models (Alpaydin 2014). In addition, the statistical theory says that averaging measurements might lead to more dependable and firmer estimates due to weaker impact of randomness (Subasi 2020). Random forest is a combination of multiple decision trees that are created from random sets of available data. Input features of a decision tree are usually randomly selected and then the predictions of the models are combined, which decreases correlation of the decision trees used and eventually increases the forest's ability to predict, to state it clearly, the accuracy of the models increase significantly (Ali et al. 2012; Alpaydin 2014). Random forests are good in managing overfitting problem, they are not so vulnerable to outliers in data, forests are easy to setup, and they find the importance and the accuracy of variables without a user (Ali et al. 2012). Random forests are simply attaining better outcomes than unique decision trees, but even so they are still interpretable (Ali et al. 2012), which is also valuable in this thesis when determining the features of the models created.

Random forests reduce correlation of the trees in a forest by not allowing single random trees to use all or not even majority of the independent variables at their splits. This approach gives other features, than the strongest predictor, more importance and makes the discrete decision trees dissimilar, which affects their intercorrelation. Therefore, random forests tend to have lower variance and are not typically suffering overfitting (James et al 2017), nevertheless, noisy datasets can cause overfitting problems (Rogel-Salazar 2018). James et al. (2017) recommends that a random forest should be set to use small number of independent variables as predictors if the training dataset contains lots of correlated independent variables.

Random forests are easy to setup. A user needs to determine the number of features in the random subset and the number of trees included within the model. Furthermore, pruning and regularization methods are possible. The correlation between the trees in a random forest should be low since the higher correlation increases the forest's error rate (Rogel-Salazar 2018).

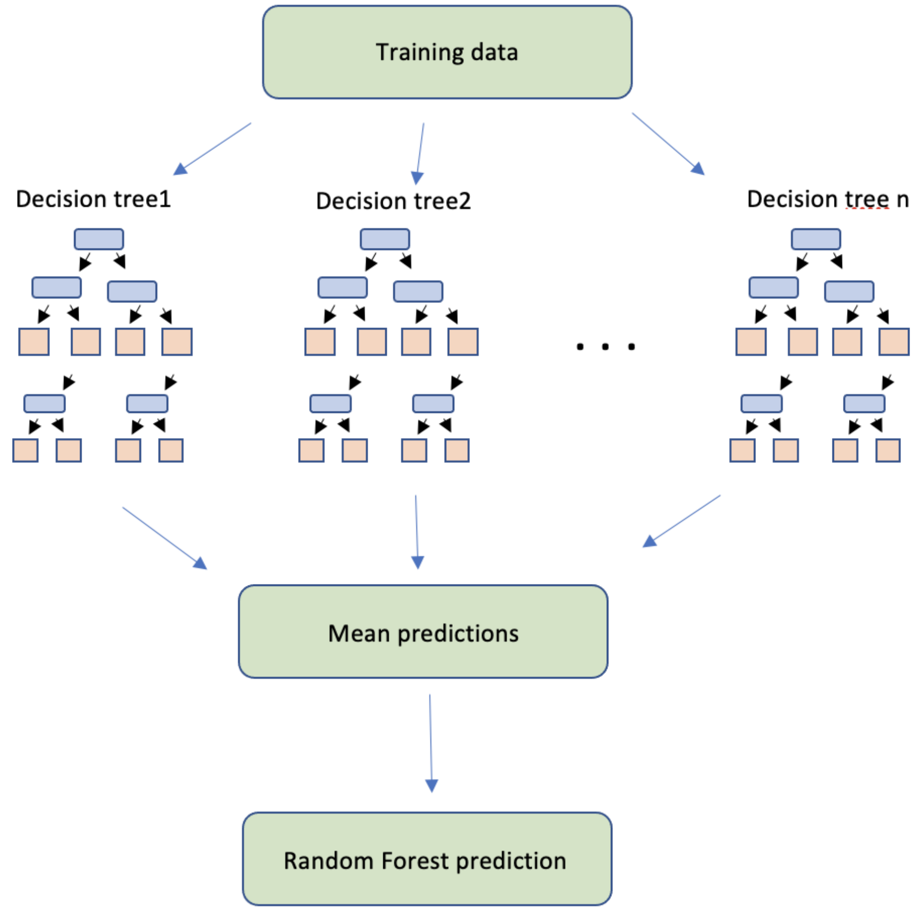

Figure 7 - Example of Random Forest Regressor

Studies show that random forest models are capable of predicting dwelling prices with robust results comparing to traditional models. The study of Jaiswal and Patil (2020) conveys that Random Forest achieves 83% R-Squared and accomplishes better than a decision tree model. Louati et al. (2022) predict dwelling prices in north of Riyadh and their random forest model outperforms decision tree and linear regression models. Hong et al. (2020) uses random forest models to predict dwelling prices of Gangnam district in Seoul, South Korea with mean percentage error of circa 5.5% while the traditional hedonic pricing model which is based on ordinary least squared linear regression achieves mean percentage error of circa 17.5%. Also, Laaksonen (2022) achieve the mean absolute percentage error of 7.83% in his master's thesis where he predicts asking prices of dwellings in Helsinki metropolitan area by using a random forest model. In addition, Zhang et al. (2022) use a random forest model for predicting prices of used dwellings in Chengdu, China and achieve a 2.16% mean absolute percentage of error. Sawant et al. (2018) used Random Forest to predict whether the closing price is below or above the listing price of a dwelling. Random Forest outperformed Decision Tree and achieved 0.999 R-Squared. Adetunji et al. (2022) use Random Forest to predict dwelling prices in Boston and achieve 0.90 R-Squared.

Extra trees method is a similar method than a random forest, but it can be a faster algorithm. It creates multiple decision trees, but samples data for each tree randomly without replacement ensuring unique data samples. Independent variables are also sampled randomly for trees. The most unique feature of the method is the features' splitting value selection, which is done randomly, also, instead of using Gini or entropy (ArcGIS Pro, n.d.). Prior studies show that the Extra Trees regression algorithm is suitable for dwelling price prediction. For instance, Mora-Garcia et al (2022) achieve R-Squared of 0.9178 with a test set. The extra trees regressor outperforms the linear regression, the random forest and the light gradient boosting machine in that study. Kumar et al. (2021) test 19 different algorithms for dwelling price prediction and the extra trees regression is the third best model. The model achieves 0.8118 R-Squared and 0.1165 mean absolute percentage error.

3.7.4 XGBoost or Extreme Gradient Boosting

Boosting is a method that improves the predictions of a single decision tree (James et al. 2017). Boosting methods use multiple decision trees to create an optimal model. In boosting, trees grow one after another, the following tree trying to learn from errors the earlier did. The trees within a boosting model are fitted on modified subsamples of the original training data. In addition, the individual trees are fitted using residuals of the moment as the response, thus, every tree tries to improve the residuals. This procedure is slow, but it improves the predictions and decreases the risk of overfitting (James et al. 2017).

Gradient boosting uses small, or weak, decision trees which are dissimilar from each other. When the model adds a new weak decision tree, the individual data observations are weighted. The observations which are already predicted well are not important in the next weak decision tree which helps the model to focus on observations which are not predicted well yet. The Gradient boosting model is an ensemble of those weak models (Korstanje 2021). The model learns slowly by using modified versions of original dataset in individual decision trees and manages the model's complexity with a regularization term, hence ultimately helping reduce overfitting (James et al. 2017; Jha et al. 2020).

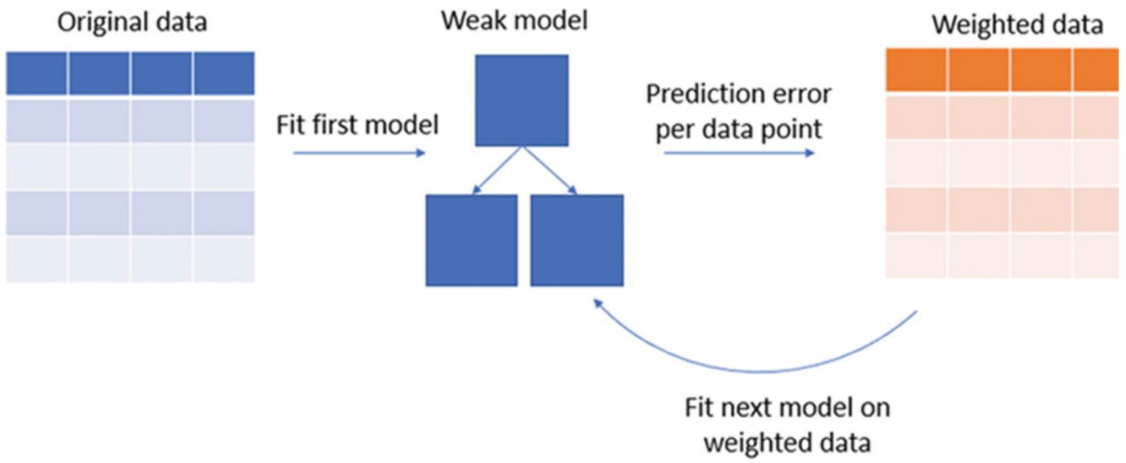
Original data
Fit first model
Weak model
Prediction error per data point
Weighted data
Fit next model on weighted data

Figure 8 - Example of the boosting process (Korstanje 2021)

The model is called gradient boosting since the model tries to optimize the gradient of the loss function. The model adds small corrections to a new weak decision tree in order to reduce the loss, hence making less errors than the earlier set of weak models did (Korstanje 2021).

Extreme Gradient Boosting alias XGBoost is a popular decision tree-based algorithm (Mansana et al. 2020). Mansana et al. (2020) argue that it is the most efficient technique for regression and classification tasks. XGBoost is an effective and scalable boosting machine learning method, which have achieved state-of-the-art results. The algorithm can run ten times faster than other existing methods due to its parallel and distributed computing and it scales to many scenarios (Chen & Guestrin 2016; Truong, et al 2019). Moreover, XGBoost models usually are more stable and have less variance (Mansana et al. 2020), which has been one of the reasons to its popularity. Decision trees need to find the best possible splits in order to make the model efficient. XGBoost does not loop through all the possible splits, due to that fact it would be time-consuming, but it uses a histogram-based splitting, which means that the XGBoost model creates histograms of every independent variable, then finds the best splits from the histograms and keeps the best total splits (Korstanje 2021).

Earlier studies show that XGBoost is able to predict dwelling prices accurately. Jha et al. (2020) suggest using XGBoost over the other models. In their study, XGBoost achieve 0.968 R-squared for the test set which is almost the same than R-Squared of a random forest model, while XGBoost outperforms the random forest model (and the other models) when MSEs are compared. XGBoost also outperforms the random forest in a study conducted by Henriksson & Werlinder (2021). XGBoost achieves MAPE 9.4% in the dwelling price prediction task with a dataset containing dwelling information from Japan (Henriksson & Werlinder). Xu and Li (2021) have similar results as XGBoost outperforms Random Forest and the other base models. In Tekin and Irem's (2022) study, XGBoost model achieves the mean absolute percentage error of 21.81%, which outperforms all the other models except the random forest model. Peng et al. (2019) used XGBoost for predicting the prices of dwellings in Southwest China and the model attained 0.9251 R-Squared. Iwai and Hamagami (2022) predict the prices per square meters of dwellings in Tokyo with XGBoost and MAE of the model is slightly above ¥ 200,000 which is circa € 1,382. The XGBoost model improves MAE by 27.3% compared to the Multiple Regression model. In the study of Yan and Zong (2020), XGBoost is the most accurate model to predict dwelling prices in Beijing among the machine learning models such as Linear Regression, Random Forest, Ridge, and Lasso models.

3.8 Which Model to Choose

As discussed in the prior chapters, there are multiple different machine learning methods and algorithms to choose. How to find the right one to the specific problem? First, it is important to analyze the goal. What is wanted to achieve? If we want to predict a specific target value, we should use supervised machine learning techniques. If the target is unclear, we should turn to unsupervised learning methods (Subasi 2020).

Supervised learning contains multiple techniques, too. Which one to choose? Again, we need to think our goal. If the target value is continuous, we can use regression techniques, while the discrete target value requires classification techniques (Subasi, 2020).

In addition, the independent variables of the dataset will affect the choosing process. However, there is no exact answer, which algorithm is the most suitable one. Therefore, the process of model choosing requires testing of different possible algorithms (Subasi 2020).

The state-of-the-art AutoML techniques can be used as well. The AutoML is an automated Machine Learning pipeline which takes care of data preprocessing, feature engineering, hyperparameter optimization and model building. It can be an efficient way to use power of Machine Learning techniques without having needed understanding of Machine Learning or how to build high-quality models (He et al. 2021).

Model selection techniques are discussed more thoroughly next.

3.8.1 Cross-Validation

The validation of a model is an important task to measure the quality of a model's inductive bias. This can be done by dividing the dataset into two sets where one is the training set and another one the validation set. The validation set is used to measure how generalized the model is. Different complexities of the model can be tested with validation sets and find the best performing model (Alpaydin 2014).

One important objective in machine learning model building is to achieve a low-test error rate, i.e., create as good predictions as possible with the test dataset. Cross-validation is a way where a subset is held out of the training dataset before fitting the model and then the model is applied to the held-out subset. A common way to use cross-validation is a k-fold cross-validation approach. In the k-fold cross-validation approach the dataset is randomly divided into k groups, typically 5 or 10 groups. The first group is the held-out subset and the rest, 1 – k, are the training set of the model. The mean squared error of the held-out group is calculated and this process is iterated k times. In the second round, the second group is the held-out group and so forth. The last step is to calculate the mean value of the calculated MSE values. Cross-validation can be used to find the actual estimate of the validation set's mean squared error. In addition, it can be used to determine the most suitable model by finding the location of minimum point of the MSE from the cross-validation curve (James et al. 2017). Typically, the k-fold cross-validation is computationally more expensive than a simpler train-test-split due its way to fit the training data multiple times, however, it usually generates better estimate of the model's performance (Brownlee 2020).

The validation set has become a part of a training data in this process, so actually there is a need of a third set, called a test set, which can be used to calculate and report the performance of the best model (Alpaydin 2014). If the dataset is divided only into two sets, the model might work well when training and test data are evaluated, but the performance with outside data is often insufficient (Huang & Le 2021 p30).

3.8.2 Hyperparameter Tuning

Machine learning models have different algorithm parameters, i.e., hyperparameters, that affect the performance of the model. The hyperparameters can be adjusted to be suitable for the task at hand. According to Garreta and Moncecchi (2013), optimal hyperparameters can have a great impact on results. One possible way to adjust different hyperparameters is just try them manually. However, that is not an optimal way since it is time-consuming.

The grid search is a better way. The Scikit-Learn has its own GridSearchCV method that can be used to find optimal hyperparameter values. A user determines which hyperparameters and which values of each hyperparameter should be tested and the GridSearchCV will test. The GridSearchCV assess all the possible combinations of hyperparameters and inputs the best performing one after evaluating them with a cross-validation (Aurelien 2019; Garreta & Moncecchi 2013). The grid search is efficient way to test different hyperparameters, but it is useful for other tasks as well, e.g., searching the best way to handle outliers and independent variables (Aurelien 2019).

Radhakrishnan's (2022) results show that hyperparameter tuning enhanced the performance of XGBRegressor from 0.9857 R-Squared to 0.9865 R-Squared. Mora-Garcia et al. (2022) results are slightly contradictory, since initial hyperparameters of Linear Regression and Random Forest models perform better than the models after the hyperparameter optimization process. However, Gradient Boosting Regressor, XGBoost, and Light Gradient Boosting Machine perform better after the hyperparameter optimization process (Garcia et al. 2022).

3.8.3 Performance Metrics

After the machine learning model is created, the next crucial step is its evaluation by using performance metrics. Model evaluation helps to figure out the effectiveness of a model in accomplishing its initial purpose. The performance metrics offer a quantitative measure how the model actually works. Usually, it is natural to use multiple distinct performance metrics which give more extensive understanding of the model's performance. The information gathered from the model evaluation can be used to select the most suitable model for that specific purpose. Next, we discuss about performance metrics which are often used in regression problems.

3.8.3.1 R-Squared

R^2 Statistic, or the coefficient of determination (Manasa et al. 2020), takes a proportion of the variance in the dependent variable, which is interpreted by the model's independent variables, in other words, it explains how well independent variables explain the variance of the dependent variable. Cause it takes a proportion the value is always between 0 and 1, where 1 means that the predicted values are perfectly explained, while 0 means that the model cannot explain the variance at all (James et al. 2017; Manasa et al. 2020). So, the aim is to create a model which R^2 is high as possible. However, the good R^2 does not guarantee that a model is good, for instance, if the test dataset is small. In some cases, it is reasonable to assume that R^2 of a model should be close to 1, especially in cases where the data are known to be linear. However, in some cases it is obvious that the model cannot explain all the variance due to its complexity and if in this kind of case the R^2 is really close to 1 or is even 1, there might be something suspicious in the model. In addition, the value of R^2 might increase when the number of independent variables increases, which is one of the metric's disadvantages (Manasa et al. 2020). Hence, it is essential to use the R^2 measure with other performance metrics. The formula for R^2 is

$$\mathrm{R}^2 = 1 - \frac{RS}{T}$$

where RS is a residual sum of squares. The formula for RS is

$$RS = \sum_{i=1}^{n} (y_i - \hat{y}\ _i)^2$$

where y is an actual dependent variable and ŷ is a predicted dependent variable. T is a total sum of squares. The formula for T is

$$T = \sum (y_i - \bar{y}_i)^2$$

The total sum of squares calculates the total amount of variance in the response dependent variable y. It represents the total variance present in the response before the actual regression attempt is performed. The residual sum of squares calculates the amount of variance which is still undefined after the regression (James et al. 2017). R² is a good metric to compare different models which can be trained with different datasets.

3.8.3.2 Adjusted R²

Such as R², adjusted R² determines the proportion of variance, a model's accuracy, in the dependent variable which is explained by the independent variables. R² tends to have an optimistic way to determine the accuracy since it increases every time when new independent variables are added in the model. Adjusted R² tries to reduce that bias. The value of an adjusted R² sets between 0 and 1, such as the value of R², but the value of adjusted R² is always less than R² (IBM 2023). The formula for adjusted R² is

$$Adjusted\ \mathrm{R}^2 = \frac{(1-\mathrm{R}^2)(\mathrm{S}-1)}{S-F-1}$$

where S is the total sample size and F the number of independent variables. Adjusted R²s' are comparable between models trained with different datasets.

3.8.3.3 MSE

MSE aka mean squared error is a performance metric that measures the amount of error in the machine learning model. It calculates the average squared disparity, in other words the average squared residual, between actual dependent variable y and predicted dependent variable ŷ. The ultimate, often impossible, goal would be to achieve 0 value for MSE. Squaring the residuals have few reasons. It extinguishes negative residuals, hence ensures that the value is either positive or zero. It also penalizes large errors more than small ones (Frost n.d.). Since the prices of dwellings are typically large numbers, the MSE is often large, too. The formula for mean squared error is

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

MSEs' of different models are not comparable if the models are trained with different datasets.

3.8.3.4 MAE

Mean absolute error (MAE) is the last performance metric discussed in this thesis. MAE calculates the average absolute error between the dependent variable y and the predicted dependent variable ŷ. The MAE is easy to understand, since it tells the actual average error in terms of the dependent variable (Allwright 2022). The formula for MAE is

$$MAE = \frac{\sum |y - \hat{y}\,|}{no}$$

where *no* is number of observations. MAEs' of different models are not comparable if the models are trained with different datasets, which is a negative side of this measure, albeit the MAE can be converted to MAPE (Mean Absolute Percentage Error) which is comparable.

3.9 Predicting

In this thesis, the author will create multiple machine learning models which try to predict prices of dwellings. One model uses independent variables X, for instance city and square meters, and try to calculate as accurate predictions ŷ as possible by creating function $\hat{f}$, that is the model's best estimate of function f, which is the exact function for predicting from independent variables X the output y, which is in this case the price of a dwelling. In other words

$$\hat{y} = \hat{f}(X).$$

Typically, only the ŷ, and especially its accuracy is the most important thing and the $\hat{f}$ is dealt as a black box. However, in this thesis the author wants to understand the models' discrepancies, and which are the most important features of the models and how y changes as a function of independent variables X. Therefore, the features of function $\hat{f}$ are crucial to understand. The models will be done for predicting purposes but also solving an inference problem (James et al. 2017). Therefore, the hyperparameters and the most important independent variables of the machine learning models will be discussed in the empirical part.

4 EMPIRICAL STUDY

This Chapter together with Chapters 2 and 3 answer to the first research question, '*How can machine learning be used to predict the price of a residential building in Finland?*' The prior chapters have created the basis of knowledge for comprehensive understanding of the dwelling markets and machine learning. This Chapter explains the practical way to create a machine learning model for the Finnish dwelling market.

As discussed earlier, the use of machine learning models has become increasingly popular in predicting dwelling prices, and Finland is no exception. In this chapter, the empirical part of the thesis is presented. The study aims to predict dwelling prices in Finland using machine learning models. Different models have been created based on three different datasets: one containing all data, one containing data gathered from Finnish cities over 100,000 residents, and the last dataset containing data gathered from Finnish cities less than 100,000 residents. The models created by using all gathered data are used as a benchmark. The purpose of this study is to compare the performance of different machine learning models trained by different type of datasets and to determine the most effective models for every dataset. In addition, the differences of the developed models will be compared.

4.1 Data

The raw dwelling transactions data used in this study were gathered from the Asuntojen.hintatiedot.fi database. The site is maintained by The Housing Finance and Development Centre of Finland. The data are supplied with cooperation agreement of the Central Federation of Finnish Real Estate Agencies. The Finnish real estate agencies Kiinteistömaailma Oy, OP Koti, Huoneistokeskus Oy, Aktia Kiinteistönvälitys Oy and RE/MAX Finland collect the data and supply them to the service. The database contains actual transaction prices and quality information of individual apartments, terraced houses and detached houses that have been sold in Finland during the previous year.

In the database, a user can search dwelling sales data of specific municipalities. The search results contain a district, a dwelling type, square meters, a debt-free price in euros, a debt-free price per square meter, a construction year, a floor level and the total floors in a building, an elevator, a condition, a plot, and an energy class information of an individual dwelling.

The district information of all individual observations is provided by real estate agents and can vary from the official district information defined by municipalities. The service provider has made some corrections in case of misspelling or if the district has multiple names, for instance, due to merged municipalities. The apartment information contains description of the rooms and special information, i.e., information about balcony, terrace, and sauna. The square meters are the living area of the dwelling. The debt-free price contains the sales price and the dwelling's share of the housing company loan at the time of sales. The construction year is the building year or the year of introduction if the building is totally renovated. The floor information describes the floor level and the total floors of the apartment house. In the case of row and detached houses, the floor information describes the number of floors in the dwelling. The elevator information shows whether the dwelling has an elevator or not. The condition information is an estimation of the dwelling's condition provided by a real estate agent or the owner of a dwelling. The plot feature is either an own plot or a rented plot. The energy class feature provides an energy efficiency rating of a dwelling. The energy efficiency rating range is from A class to G class (ARA n.d.). The data were collected between March 2022 and October 2022 and include transactions that took place between March 2021 to October 2022.

Spatial data affect highly on dwelling prices as discussed earlier; therefore, additional data of districts' characteristics were needed. Data of characteristics of districts were collected from the Paavo database provided by Statistics Finland. The database contains information on the population structure, education levels, income levels, housing types, workplaces, households' life stages and residents' main activities, for instance, the proportion of students, employed, and unemployed individuals in the district. The information is arranged based on the postal codes used by the Finnish postal service (Statistics Finland n.d.). The Paavo data used in this study were published in January 2022.

As mentioned above, the data contain information by postal code areas. The population structure data contain information on total inhabitants, the number of males and females, the average age, and the different age bins in 2020. The educational structure data have information on the number of residents aged 18 or over, and the type of educations completed by residents, e.g., basic level studies, matriculation examination, vocational diploma, lower and upper university degree. The residents' disposable monetary income feature contains information of residents' mean and median income, accumulated purchasing power, number of residents in the lowest, middle-, and highest-income category. The income data are created from tax information. The size and stage in life of household data describe the total number, average size, and the type of households, e.g., one-person house, pensioner households, and households with small children. In addition, that data recount the areas' occupancy rates and the number of households living in rented and own dwellings (Statistics Finland, n.d.).

The households' disposable monetary income has same information than inhabitants' disposable monetary income, but the information is converted into household level. The buildings and dwellings data describe the total number of buildings, and the number of residential buildings, free-time buildings, and other type of buildings. Moreover, it contains information of average floor area. The workplace structure data have information on total number of workplaces, number of primary, processing, service, agriculture, forestry, fishing, mining, manufacturing, wholesale, retail trade and other types of workplaces. The main type of activity data tells the number of employed, unemployed, children, students, and pensioners in the area (Statistics Finland, n.d.).

The transaction data for this study from Asuntojen.hintatiedot.fi were gathered by using the ParseHub web scraper. The ParseHub automates the web scraping process. The ParseHub users do not need to know coding but are still able to extract important information from websites.

The raw transaction data gathered by the ParseHub contain 19.800 rows from 19 Finnish cities. The raw data contained lots of duplicate rows and after removing the duplicates the dataset contains 11.570 rows. Every row has 13 features. The 19 cities used in this study are Espoo, Hämeenlinna, Joensuu, Jyväskylä, Järvenpää, Kokkola, Kouvola, Kuopio, Lahti, Lappeenranta, Mikkeli, Nurmijärvi, Oulu, Pietarsaari, Pori, Savonlinna, Seinäjoki, Tampere, and Turku. Helsinki is the largest city in Finland but excluded from the dataset since it has been used in almost every similar study conducted in Finland and the author wanted to do the research from a different perspective.

City	Observations	Mean price	Mean square meters
Espoo	2376	343,366.39	80.54
Tampere	1471	205,894.22	60.99
Turku	1155	189,359.29	67.37
Oulu	1005	148,642.03	67.08
Lahti	967	138,535.07	75.94
Jyväskylä	860	145,760.58	65.67
Kuopio	810	152,863.84	72.64
Joensuu	494	136,691.98	70.16
Hämeenlinna	354	131,418.85	71.37
Pori	341	103,987.94	72.35
Järvenpää	282	199,191.07	71.00
Kouvola	280	93,977.70	85.90
Nurmijärvi	269	238,194.63	97.60
Lappeenranta	267	142,320.64	71.13
Seinäjoki	185	130,457.11	67.70
Kokkola	155	136,971.13	79.59
Mikkeli	128	101,723.40	65.82
Savonlinna	103	60,939.44	62.86
Pietarsaari	68	121,543.99	72.82

Table 1 - Basic statistics of observations in each city

Table 1 describes the transaction observations by the city level. Espoo has most observations while Pietarsaari has least. The highest price mean value is in Espoo and the lowest in Savonlinna. The observations from Nurmijärvi have the largest mean living space and the smallest mean living space is in Tampere.

Idea of this study is to compare machine learning models that are trained with different sets of data gathered from different cities. In this study, Espoo, Jyväskylä, Kuopio, Lahti, Oulu, Tampere, and Turku are considered as large cities and the rest of the cities are considered as small cities. The large cities contain all Finnish cities that have over 100,000 residents (excluding Helsinki and Vantaa), which are also among the growth centers of Finland (KTI 2023). The rest of the cities have less than 100,000 residents (Hagerlund 2022).

Variable	Non nulls	Nulls	Type	Description	Mean
Kaupunki	11570	0	object	City	-
Area	11570	0	object	District in a city	-
Huoneisto	11560	10	object	Description of the dwelling	-
Type	11567	3	object	Type of the dwelling (block house, row house etc.)	-
Sm2	11567	3	float64	Square meters	72.05
Price	11567	3	float64	Price	196771.88
Em2	3747	7823	float64	Price per square meter	3917.64
Year	11569	1	float64	Building year	1987.76
Floor	10713	857	object	Floor level / total floors	-
Elevator	11567	3	object	Dwelling has an elevator	-
Condition	10157	1413	object	Condition	-
Plot	11366	204	object	Plot type	-
Energyclass	10115	1455	object	Energy class	-

Table 2 - Description of columns in the raw transactions data

Table 2 shows the columns of the raw transactions data. The data contain a lot of categorical features that needs to be converted into more suitable forms for machine learning models. The Price variable is the dependent variable, and the rest of the variables are independent variables in this study. The mean price is EUR 196,771.88, the mean square meters 72.05, and the mean building year 1987.76. These values appear reasonable. The raw transactions data included missing values that need closer examination.

Table 3 describes statistics of the numerical values of the raw transaction dataset. It is evident that the dataset contains some incorrect values.

Variable	Count	Mean	Standard deviation	Min	Max
Square meters	11,567	72.05	37.95	1	1,277
Price	11,567	196,772	147,787	1	3,400,000
Year	11,569	1988	23,30	1,800	2,024

Table 3 - Statistics of numerical features

The square meters minimum value is 1 m2 which is smaller than a proper dwelling would be. In addition, the minimum price is EUR 1, which is probably a mistake. The maximum construction year 2024 indicates that there have been sold a new dwelling which is under construction. Some of the variables in the raw transactions data are nonnumeric or are in nonnumeric form and need feature engineering measurements.

Variable	Type	Description	Mean
Postinumeroalue	object	District with postal code	-
Asukkaat yhteensä, 2020 (HE)	int64	Total number of residents	3225.81
Asukkaiden keski-ikä, 2020 (HE)	int64	Average age of residents	43.52
Asukkaiden keskitulot, 2020 (HR)	int64	Average income of residents	24720.9
Asukkaiden mediaanitulot, 2020 (HR)	int64	Median income of residents	21778.56
Asukkaiden ostovoimakertymä, 2020 (HR)	int64	Cumulative purchasing power of residents	67982627.93
Taloudet yhteensä, 2020 (TE)	int64	Total number of households	1683.43
Talouksien keskikoko, 2020 (TE)	float64	Average size of households	1.97
Talouksien keskitulot, 2020 (TR)	int64	Average income of households	40876.85
Talouksien mediaanitulot, 2020 (TR)	int64	Median income of households	34617.65
Kesämökit yhteensä, 2020 (RA)	int64	Total cottages in the area	134.25
Rakennukset yhteensä, 2020 (RA)	int64	Total buildings in the area	665.25
Asunnot, 2020 (RA)	int64	Dwellings in the area	1864.05
Asuntojen keskipinta-ala, 2020 (RA)	float64	Average sqm of dwellings in the area	88.12
Työpaikat yhteensä, 2019 (TP)	int64	Total number of jobs in the area	1337.11
Asukkaat yhteensä, 2019 (PT)	int64	Total number of residents in the area	3204.46
Työlliset,%	float64	Proportion of employed to total residents in the area	0.41
Työttömät,%	float64	Proportion of unemployed to total residents in the area	0.05
Lapset%	float64	Proportion of children to total residents in the area	0.15
Opiskelijat,%	float64	Proportion of students to total residents in the area	0.07
Eläkeläiset,%	float64	Proportion of retired to total residents in the area	0.28
Muut,%	float64	Proportion of unclassified people to total residents in the area	0.03

Alkutuotannon%	float64	Proportion of primary production jobs to total jobs in the area	0.12
Jalostuksen%	float64	Proportion of secondary production/processing jobs to total jobs in the area	0.22
Palveluiden%	float64	Proportion of service industry jobs to total jobs in the area	0.6
Pientaloasunnot,%	float64	Proportion of one-family houses to total dwellings in the area	0.69
Kerrostaloasunnot,%	float64	Proportion of flats to total dwellings in the area	0.29
Muut%	float64	Proportion of other dwellings to total dwellings in the area	0.15
Koulutetut%	float64	Proportion of residents with some education to all over 18 years in the area	0.75
Perusasteen%	float64	Proportion of residents with basic education to all over 18 years in the area	0.22
Ylioppilastutkinnon%	float64	Proportion of secondary school graduates to all over 18 years in the area	0.05
Ammatillisen%	float64	Proportion of vocational school graduates to all over 18 years in the area	0.49
Alemman%	float64	Proportion of residents with a bachelor’s degree to all over 18 years in the area	0.12
Ylemmän%	float64	Proportion of residents with master’s degree to all over 18 years in the area	0.1
Miehet,%	float64	Proportion of men to total residents in the area	0.5
Naiset,%	float64	Proportion of women to total residents in the area	0.48
Omistusasunnoissa%	float64	Proportion of households living in an own dwelling to total households in the area	0.69
Vuokra-asunnoissa%	float64	Proportion of households living in a rented dwelling to total households in the area	0.25
Muissa%	float64	Proportion of households that do not own or rent their dwelling to total households in the area	0.02
Asuinrakennukset%	float64	Proportion of dwellings to total buildings in the area	0.85
18 täyttäneet	float64	Proportion of residents over 18 years old to total residents in the area	0.82

Table 4 - Description of spatial dataset

The spatial dataset contained mostly amounts of specific features in a district. The absolute numbers are not relevant on every occasion since some districts might have much more residents compared to some others; therefore, the total amounts might give distorted information. The share of specific features, such as proportion of employed residents or proportion of residents holding a master’s degree, have more importance than absolute numbers when comparing districts. Therefore, feature engineering measures were conducted in pre-processing phase in order to produce more informative independent variables. Table 4 explains the features of districts dataset and shows the mean values of features.

The raw transaction data did not contain postal codes and the district definitions were not the same ones than in the spatial data collected from the Paavo database. Therefore, the datasets were in a form that they were impossible to merge. The author wrote a web scraper code with Python that retrieved postal codes for every district available in the raw transaction dataset. The web scraper took a city and district information from the raw transaction dataset and feed the data into the Google Maps search. Then, the web scraper retrieved the postal code in question from the Google Maps search results and created a list that contained the postal codes for districts. Final step was to concatenate the postal codes to the raw transaction data and merge the raw transaction and the spatial datasets.

As discussed in Chapter 2, loan levels and interest rates have influence on dwelling prices. Loan features and interest rate features were left out from this thesis since the sales data, or the spatial data did not contain information regarding those features. The use of benchmark rates, such as the Euribor 12-months, as a feature, would have been possible. However, due to fact that the sales data did not contain exact dates for transactions, the use of a benchmark rate, which fluctuates every day, as a feature, would have led to misleading results, therefore, the author did not use it.

4.2 Missing values

Missing values can be handled multiple ways as previously discussed. Figure 9 shows that some of the features of the transaction dataset have missing values. The feature Em2 describes the price per square meter of a dwelling which is not a relevant independent variable in the scope of this study. i.e., the aim of the study is to predict the price, therefore, the price per square meter is not known in the unseen new data. For that reason, the Em2 variable is ignored. The rest of the independent variables that contain missing values are relevant and they require data pre-processing measures.

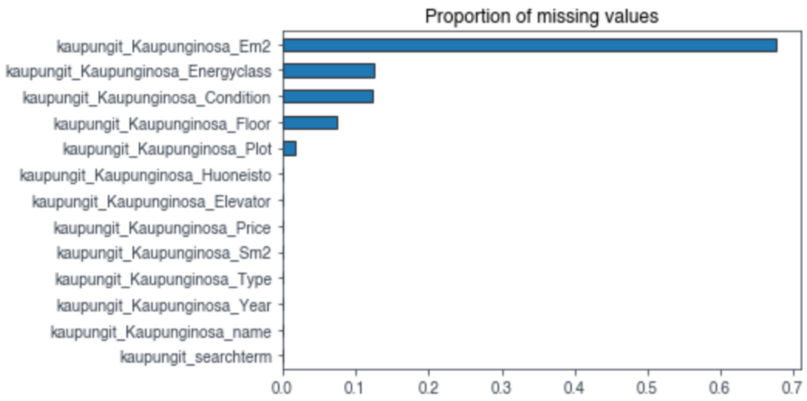

Figure 9 – Proportion of missing values in raw transaction data

The Price variable contained 3 missing values (0.02%). All the observations without the price feature were removed. Both, the Condition and the Energy class variables have more than 10% (circa 12.5% each) missing of the total which can lead to a change of the variable's distribution when conducting a mean imputation as Jadhav et al. (2019) described. However, both independent variables are important for this study and the dataset is relatively small, hence, the author decided to keep them and fill the missing values with mean values. The missing values of the condition variable filled with a value 'tyyd' which stands for satisfactory condition. The missing values of the energy class variable filled with a value 'ok_e' which stands for satisfactory energy class.

The Floor and the Plot variables had 7.4% and 1.8% of the values missing, respectively. The Floor variable was first divided into two variables: the floor level of the dwelling and the total floors in the building. Then, the missing values of both floor variables were replaced with the mean values. The missing values of the plot variable filled with a value 'unknown'.

The Huoneisto variable had total of 10 (0.08%) missing values which were replaced with a value 'unknown'. The Year had 1 (0.009%) value missing which was replaced with a value 1990.

4.3 Outliers

Good machine learning models are well generalized. Outliers, which are abnormal observations in the dataset, might mislead the model to learn distorted patterns. Different cities can have different characteristics in terms of dwellings; therefore, the outliers were screened by cities individually. Various reasons can affect occurrence of outliers. In this study, the data were gathered from a database where real estate agents were imported data. Some of the inputs might be mistakes due to human error.

Even though, the cities have different characteristics, the author decided to remove all observations that had the price less than EUR 10,200 or the price over EUR 1,400,000, prior to further analysis. The prices less than EUR 10,200 and over than EUR 1,400,000 can be real observations, and in the expensive cases probably are, but those prices do not present general dwelling prices in Finnish dwelling market and therefore excluded. Same measures were conducted to the feature of square meters. The author decided to keep the observations between 16m2 – 245m2 before more extensive analysis.

The figure 10 shows that the price outliers in some cities are not necessarily outliers in some other cities. Therefore, it is necessary to screen the cities individually. For instance, the prices over circa EUR 400,000 are outliers in Järvenpää while they are normal in Espoo. The matplotlib library's boxplot was used to detect outliers. The boxplot calculates first and third quartiles from the data and creates a box. Then, it extends whiskers from the box by one and a half times the inter-quartile range. All the observations outside the whiskers are considered as outlier (Matplotlib n.d.).

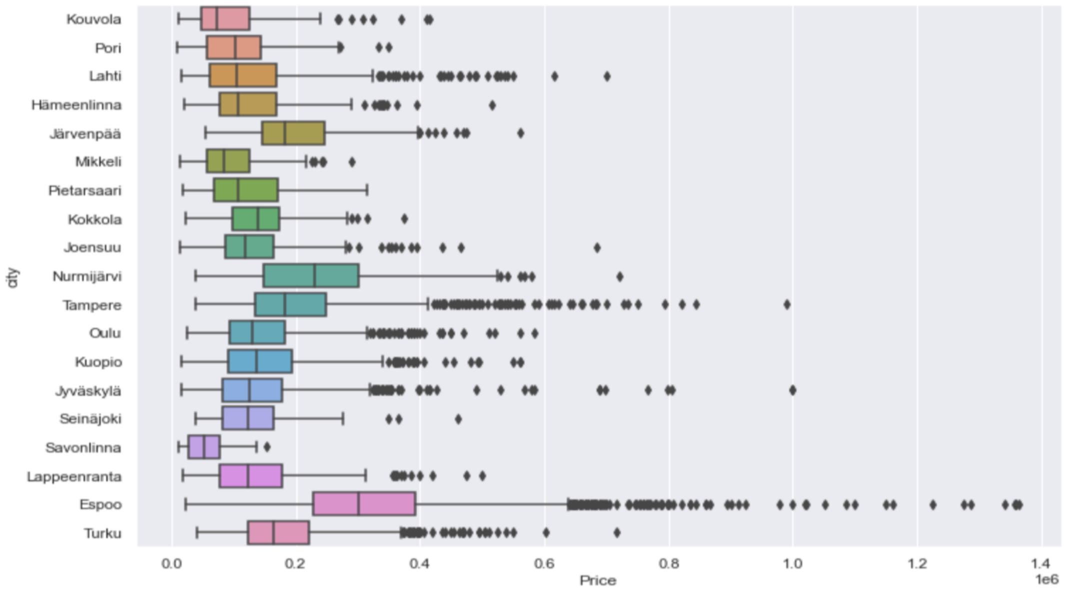

Figure 10 - Price outliers

Osborne & Overbay (2004) suggested that variables can be transformed into logarithmic scale in order to decrease the number of outliers. Figure 11 presents that the number of outliers is significantly lower after the variable is transformed into natural logarithmic scale. However, outliers are still present, and they need measures. The logarithm transformation is discussed more in the next chapter.

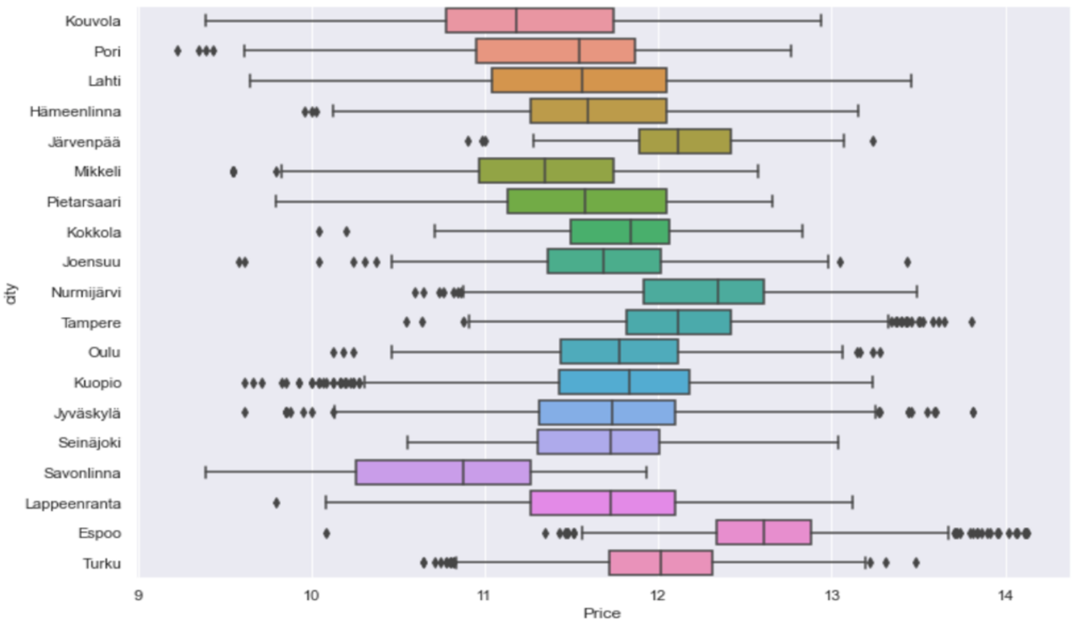

Figure 11 - Price outliers after log transformation

The author decided to remove all outlier values independently. Figure 12 shows the result.

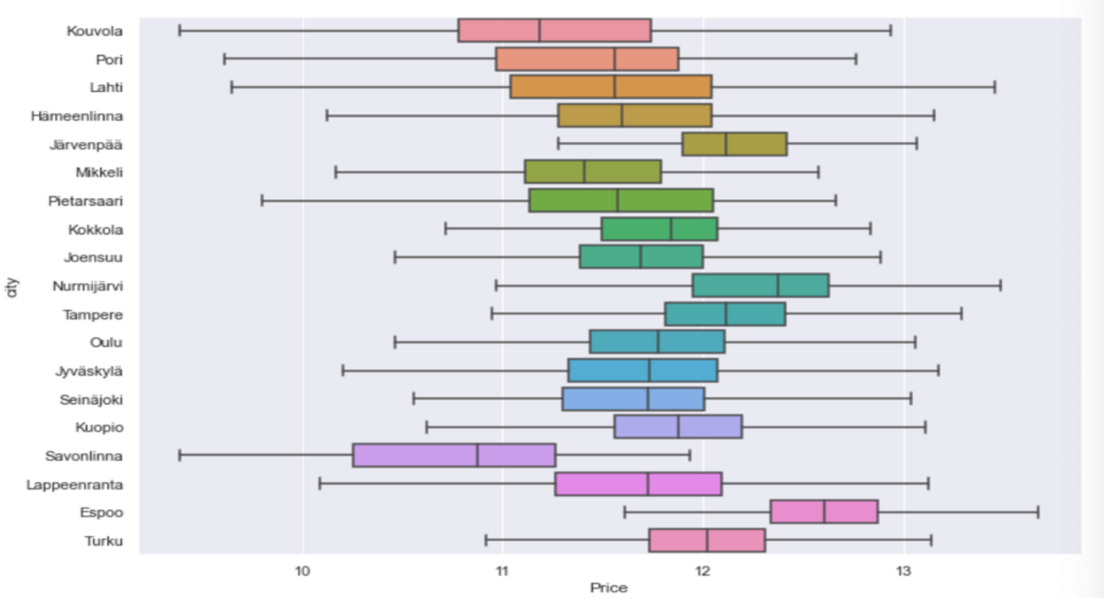

Figure 12 - Prices after outliers were removed

The square meters of dwellings variable had same issues with outliers, though, the distributions of square meters were more equal than the distributions of prices.

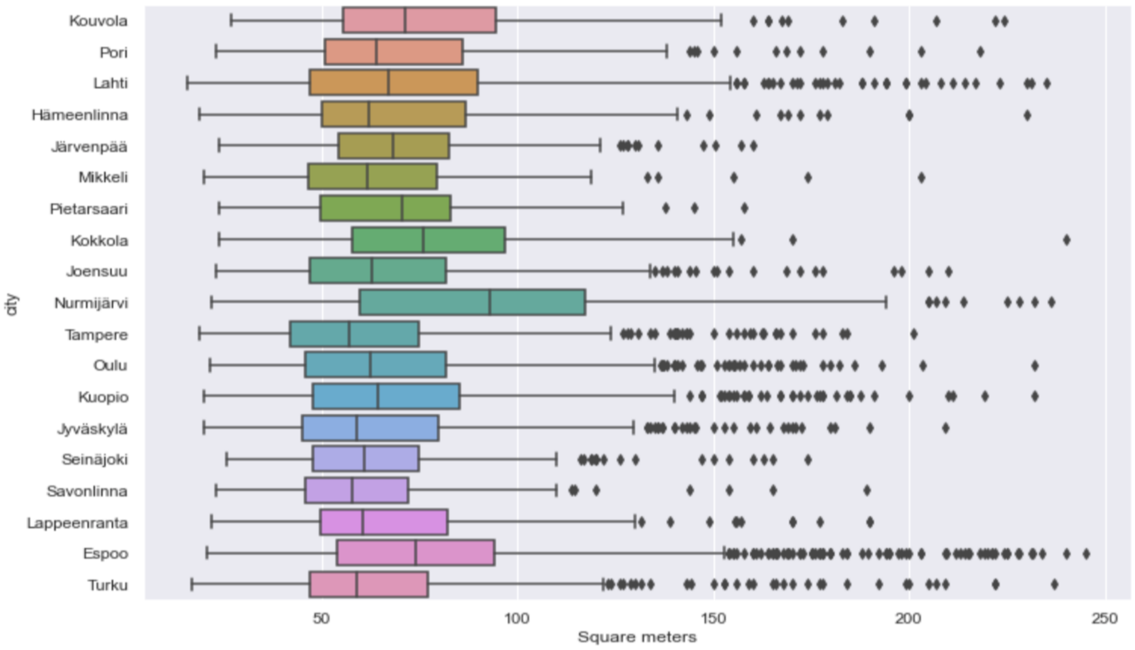

Figure 13 - Square meter outliers

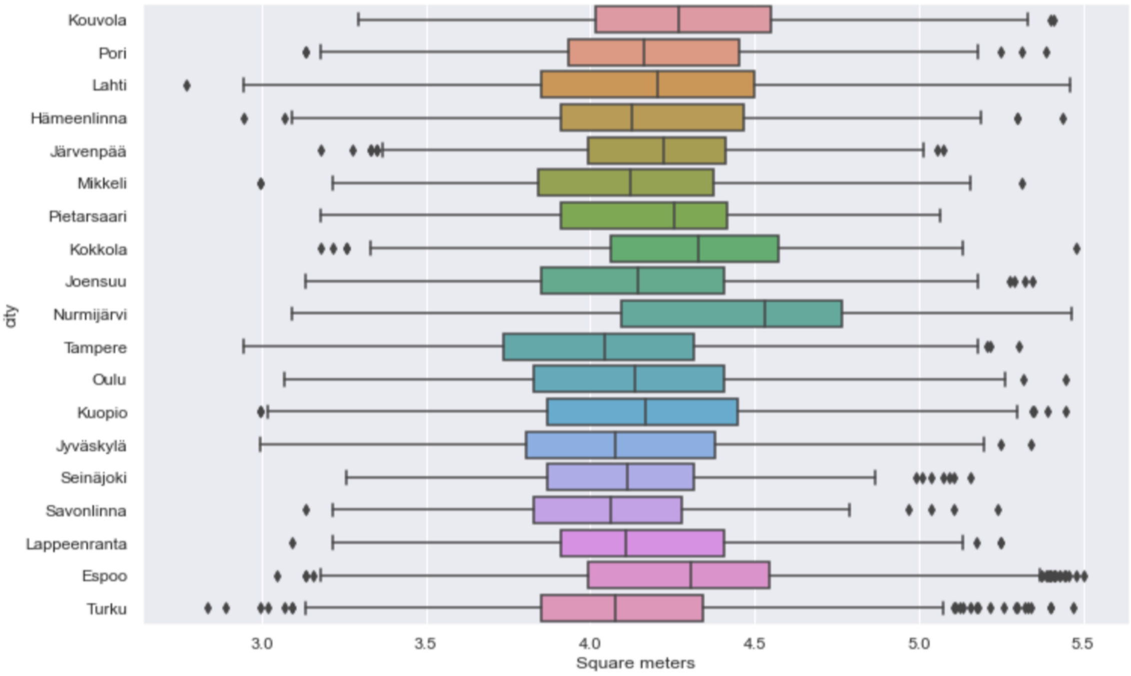

Figure 14 - Square meters of the dwellings after log transformation and before outliers' removal

As seen in Figure 14, the number of outliers is significantly decreased after natural logarithmic transformation. Nevertheless, the author decided to remove all observations containing square meter outliers in order to enhance the generalization of the model. Figure 15 presents the square meters variable after the removal of outliers.

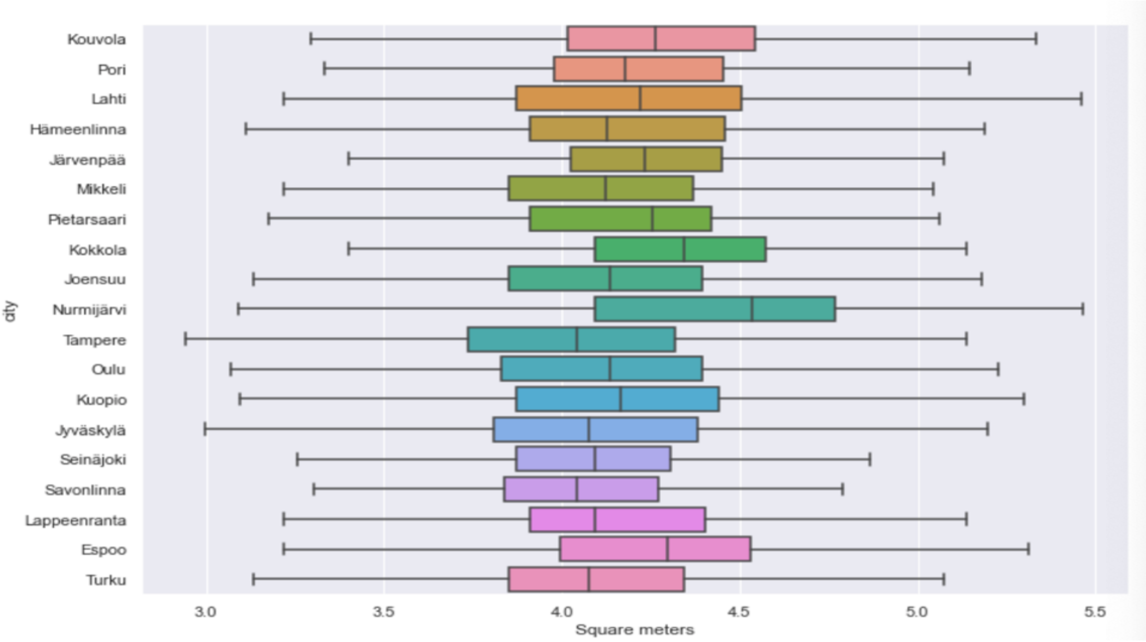

Figure 15 - Square meters of the dwellings after log transformation and outliers' removal

4.4 Log transformation

For comparing purposes, the first benchmark model was developed with the whole data containing all cities of this study. Then, the large cities model and small cities model were created. Since the datasets contained different cities with different characteristics, the distributions of the datasets were screened independently.

Some of the variables, such as the dependent variable price and the independent variable square meters of the dwelling had skewed distributions. Tree based models can handle distributions that are not normally distributed, but some other models may struggle with that kind of data. In this study, the price and the square meters of the dwelling variables were log transformed to solve the issue of skewed distributions. Next, the log transformations of all datasets are visualized.

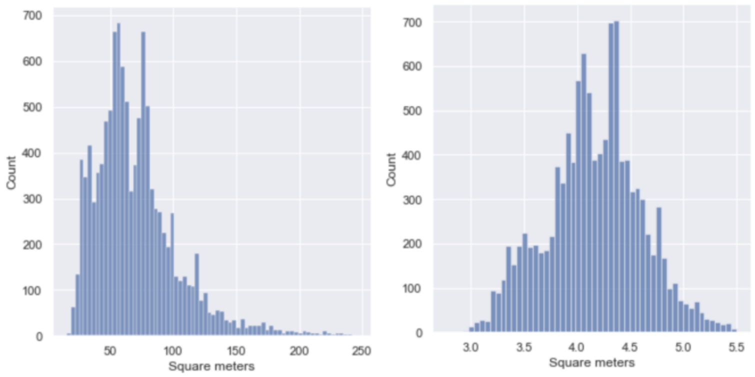

Figure 16 – The square meters of the dwellings in the whole dataset before and after the log transformation

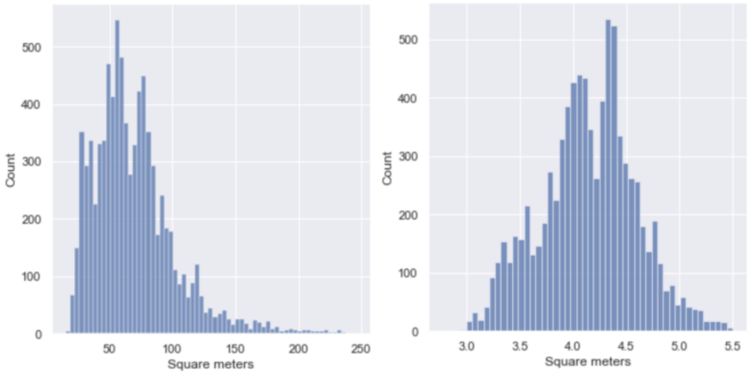

Figure 17 – The square meters of the dwellings in the large cities' dataset before and after the log transformation

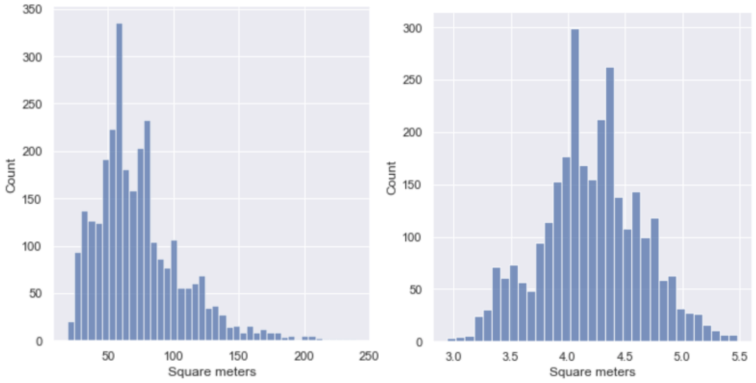

Figure 18 – The square meters of the dwellings in the small cities' dataset before and after the log transformation

The square meters of dwellings variable are in all datasets positively skewed, which means that most of the observations are closer the left tail of the distribution, but the right tail is longer and have fewer observations. After the natural logarithm transformation, the distributions are significantly more normally distributed, while most of the observations are around median and the tails are nearly steady.

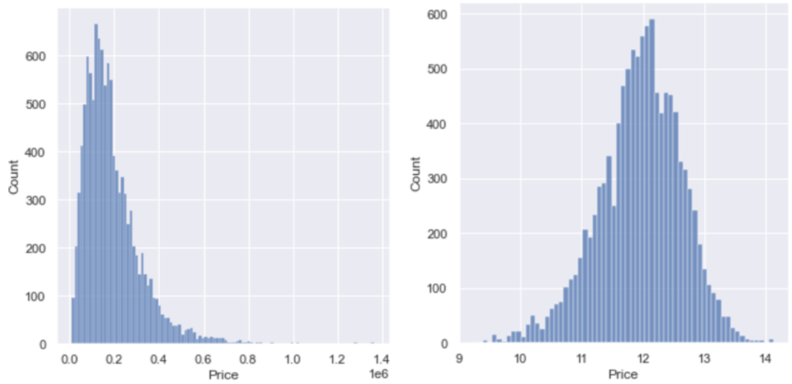

Figure 19 -The Prices of the dwellings in the whole dataset before and after the log transformation

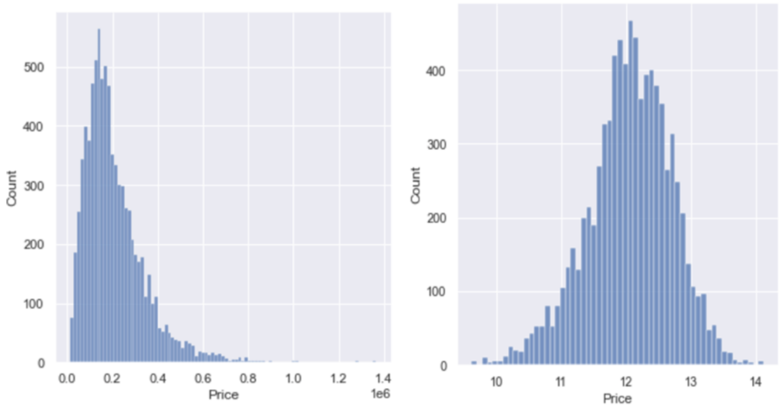

Figure 20 – The prices of the dwellings in the large cities' dataset before and after the log transformation

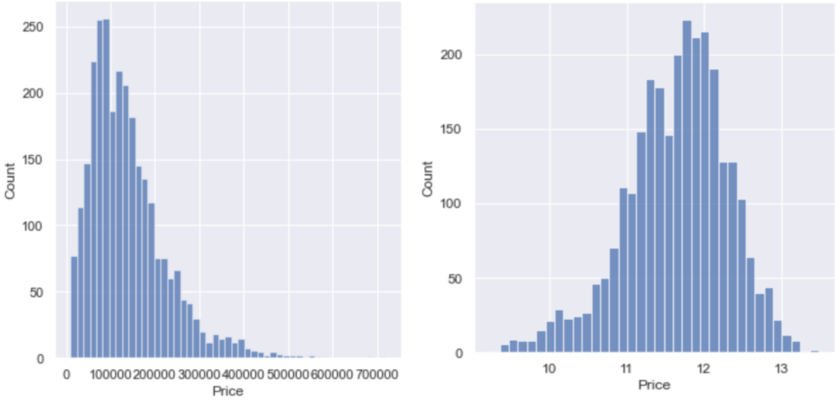

Figure 21 – The prices of the dwellings in the small cities' dataset before and after the log transformation

The price variable has same phenomenon. The prices are more normally distributed after the natural logarithm transformation.

4.5 Feature Engineering

The raw transactions dataset contained various variables that did not have numerical information, or the information was in nonnumerical form. That can be problematic for machine learning models and often variables are useless in that form. For instance, the Huoneisto variable contained information on the characteristics of a dwelling, e.g., how many rooms it has, whether it has a balcony, a sauna, a yard, or a terrace. The information was presented in text strings and was therefore useless in that form. The information on the rows of the Huoneisto variable were divided into multiple variables, which contained the original variable's information. The new variables were the number of rooms, the yard and terrace, and the balcony variables.

The Floor variable contained the actual floor level of the dwelling and the total number of floors in the building; therefore, the variable was sliced into two parts similarly than the Huoneisto variable was divided into multiple parts.

4.5.1 One-hot encoding & bins

One-hot encoding is a solution which converts categorical variables into multiple binary variables. One-hot encoding was used to create dummy variables from the Kaupungit, the Energy class, the Year, the Condition, the Plot, and the Type variables.
The energy class and the year variables contained many different values; therefore, the data of those variables were transformed into bins that described information better.

One way to handle noisy data is transforming the data into bins, which means that continuous data are divided into specific categories. The year independent variable was transformed into five categories based on the building years of dwellings. The used intervals of the bins were 0, 1959, 1979, 1999, 2018, and 2050. The energy class variable was transformed into 3 bins, which were bad_e, ok_e, and good_e, based on the dwelling’s energy classes. After the binning process, the variable was one-hot encoded, and the oldest and the poorest category were ignored from the dataset to avoid the dummy variable trap, which means situation where the dummy variables have high correlation between each other.

4.6 Feature Selection

Feature selection is mainly done to prevent overfitting problem. Some of the variables had high correlations between themselves, therefore only one of them included into the final datasets used in modelling. Following columns were removed due to high correlation or multicollinearity after the first screening:

Area
Floors total
elevator_on
5h
6h
7h
8h
Alkutuotannon%
Jalostuksen%
Kerrostaloasunnot,%
Miehet,%
postinumero
Asukkaiden keskitulot, 2020 (HR)
Talouksien keskitulot, 2020 (TR)
Talouksien mediaanitulot, 2020 (TR)
unknown

las.parveke
18 täyttäneet
Lapset%
Työttömät,%
Opiskelijat
Eläkeläiset,%
Muut,%
Asukkaat yhteensä, 2019 (PT)
Koulutetut%
Perusasteen%
Ylioppilastutkinnon%
Ammatillisen%
Alemman%
Vuokra-asunnoissa%
Muissa%
Taloudet yhteensä, 2020

Table 5 – The removed independent variables

Variable	Type	Mean	Min	Max	Standard deviation
Sm2	float64	70.6	16.0	245.0	32.9
Price	float64	193014.22	10200.0	1400000.	133787.17
ok	int64	0.09	0.0	1.0	0.29
rt	int64	0.24	0.0	1.0	0.43
vuokra	int64	0.25	0.0	1.0	0.43
hyvä	int64	0.66	0.0	1.0	0.47
tyyd	int64	0.32	0.0	1.0	0.47
good_e	int64	0.32	0.0	1.0	0.47
ok_e	int64	0.48	0.0	1.0	0.5
1 h	float64	0.14	0.0	1.0	0.35
2 h	float64	0.38	0.0	1.0	0.48
3 h	float64	0.3	0.0	1.0	0.46
4 h	float64	0.17	0.0	1.0	0.37
piha/terassi	float64	0.0	0.0	1.0	0.06
parveke	float64	0.11	0.0	1.0	0.32
Floor	int64	2.44	1.0	23.0	1.81
Year_Very old	int64	0.32	0.0	1.0	0.47
Year_old	int64	0.24	0.0	1.0	0.43
Year_fresh	int64	0.23	0.0	1.0	0.42
Year_new	int64	0.13	0.0	1.0	0.33
Asukkaat yhteensä, 2020 (HE)	float64	9384.12	53.0	28449.0	6105.15
Asukkaiden keski-ikä, 2020 (HE)	float64	41.4	27.0	57.0	4.23
Asukkaiden mediaanitulot, 2020 (HR)	float64	23046.58	15967.0	40243.0	3867.11
Asukkaiden ostovoimakertymä, 2020 (HR)	float64	20496482	126220	6802490	138573786.6
Talouksien keskikoko, 2020 (TE)	float64	1.89	1.3	3.3	0.36
Kesämökit yhteensä, 2020 (RA)	float64	65.25	0.0	1875.0	138.51
Rakennukset yhteensä, 2020 (RA)	float64	1155.93	17.0	3889.0	671.47
Asunnot, 2020 (RA)	float64	5812.93	39.0	22290.0	4418.86
Asuntojen keskipinta-ala, 2020 (RA)	float64	72.54	46.1	175.5	16.29
Työpaikat yhteensä, 2019 (TP)	float64	4809.22	1.0	29311.0	6482.96
Työlliset,%	float64	0.44	0.25	0.57	0.05
Palveluiden%	float64	0.8	0.0	1.0	0.15
Pientaloasunnot,%	float64	0.36	0.0	1.0	0.29
Muut%	float64	0.16	0.02	0.92	0.11
Ylemmän%	float64	0.16	0.01	0.44	0.08
Naiset,%	float64	0.52	0.42	0.58	0.02
Omistusasunnoissa%	float64	0.54	0.05	0.97	0.16
Asuinrakennukset%	float64	0.84	0.08	0.98	0.11
kaupunki_Espoo	uint8	0.2	0.0	1.0	0.4
kaupunki_Hämeenlinna	uint8	0.03	0.0	1.0	0.17
kaupunki_Joensuu	uint8	0.04	0.0	1.0	0.2
kaupunki_Jyväskylä	uint8	0.08	0.0	1.0	0.27
kaupunki_Järvenpää	uint8	0.02	0.0	1.0	0.16
kaupunki_Kokkola	uint8	0.01	0.0	1.0	0.12
kaupunki_Kouvola	uint8	0.02	0.0	1.0	0.15
kaupunki_Kuopio	uint8	0.07	0.0	1.0	0.26
kaupunki_Lahti	uint8	0.08	0.0	1.0	0.27
kaupunki_Lappeenranta	uint8	0.02	0.0	1.0	0.15
kaupunki_Mikkeli	uint8	0.01	0.0	1.0	0.11
kaupunki_Nurmijärvi	uint8	0.02	0.0	1.0	0.15
kaupunki_Oulu	uint8	0.09	0.0	1.0	0.28
kaupunki_Pietarsaari	uint8	0.01	0.0	1.0	0.08
kaupunki_Pori	uint8	0.03	0.0	1.0	0.17
kaupunki_Savonlinna	uint8	0.01	0.0	1.0	0.09
kaupunki_Seinäjoki	uint8	0.02	0.0	1.0	0.13
kaupunki_Tampere	uint8	0.13	0.0	1.0	0.34
kaupunki_Turku	uint8	0.1	0.0	1.0	0.3

Table 6 – Variables after the feature engineering

Table 6 presents the variables after the feature engineering and the first feature selection.

Next, the p-values of independent variables were assessed to determine which variables have predicting power. The following table shows the variables which p-values were higher than 0.05. Those variables were not statistically significant; hence, the author removed them. It is important to notice that the p-values were different in different datasets, hence, different independent variables were removed.

Removed from the whole dataset	**p-value**	**Removed from the large cities' dataset**	**p-value**	**Removed from the small cities' dataset**	**p-value**
ok_e	0.68	good_e	0.088	ok_e	0.645
piha/terassi	0.499	ok_e	0.432	piha/terassi	0.53
Asukkaat yhteensä, 2020 (HE)	0.196	piha/terassi	0.473	parveke	0.142
Asukkaiden keski-ikä, 2020 (HE)	0.053	Talouksien keskikoko, 2020 (TE)	0.758	Asukkaat yhteensä, 2020 (HE)	0.424
Talouksien keskikoko, 2020 (TE)	0.585	Kesämökit yhteensä, 2020 (RA)	0.079	Asukkaiden keski-ikä, 2020 (HE)	0.155
Rakennukset yhteensä, 2020 (RA)	0.054	Rakennukset yhteensä, 2020	0.326	Asukkaiden mediaanitulot, 2020 (HR)	0.457
Työpaikat yhteensä, 2019 (TP)	0.892	Palveluiden%	0.626	Asukkaiden ostovoimakertymä, 2020 (HR)	0.128
Palveluiden%	0.411			Talouksien keskikoko, 2020 (TE)	0.499
Omistusasunnoissa%	0.491			Rakennukset yhteensä, 2020 (RA)	0.494
kaupunki_Mikkeli	0.932			Asunnot, 2020 (RA)	0.191
				Asuntojen keskipinta-ala, 2020 (RA)	0.436
				Palveluiden%	0.432
				Pientaloasunnot, %	0.102
				Omistusasunnoissa%	0.833

Table 7 - The independent variables of the whole, large cities, and small cities dataset with p-value over 0.05

After the removal of the independent variables shown in Table 5 and 7, some variables still had high correlations between each other, therefore, further cleaning was needed. The independent variables that had over 50% absolute correlation with each other were screened and some of them were removed in order to reduce the complexity of the models. After the removal, the absolute correlation between the independent variables were below 50%. The removed independent variables are shown in Table 8.

Whole	Large cities	Small cities
tyyd	Muut%	Muut%
Asunnot, 2020 (RA)	tyyd	tyyd
Asukkaiden mediaanitulot, 2020 (HR)	Asukkaat yhteensä, 2020 (HE)	Asuinrakennukset%
1 h	Asuntojen keskipinta-ala, 2020 (RA)	1 h
Pientaloasunnot, %	Omistusasunnoissa%	
Muut%	Asukkaiden mediaanitulot, 2020 (HR)	
Asuinrakennukset%	Työpaikat yhteensä, 2019 (TP)	
	1 h	
	Asunnot, 2020 (RA)	
	Asuinrakennukset%	

Table 8 - The removed independent variables by datasets

4.7 Scaling

Tables 8, 9, and 10 show the final datasets before scaling measurements. The tables show that the independent variables were in different scales after the earlier pre-processing steps. That can cause problems as discussed earlier, since machine learning models can overvalue variables that have higher values compared to other variables, even though that is unrelated to the predicting power of the variables.

Variable	Type	Mean	Description
Sqm (logarithm)	float64	4.16	Square meters
Price (logarithm)	float64	11.96	Price
rt	int64	0.25	Row house
vuokra	int64	0.25	Rental plot
hyvä	int64	0.67	Good condition
good_e	int64	0.32	Good energy class
2 h	float64	0.38	2 rooms
3 h	float64	0.3	3 rooms
4 h	float64	0.17	4 rooms
parveke	float64	0.11	Has a balcony
Floor	int64	2.44	Floor level
Year_Very old	int64	0.31	Building year between 1960 - 1979
Year_old	int64	0.24	Building year between 1980 - 1999
Year_fresh	int64	0.23	Building year between 2000 - 2018
Year_new	int64	0.13	Building year 2019 -
Asukkaiden ostovoimakertymä, 2020 (HR)	float64	206415088.29	Cumulative purchasing power of residents
Kesämökit yhteensä, 2020 (RA)	float64	64.2	Number of cottages in the district
Asuntojen keskipinta-ala, 2020 (RA)	float64	72.45	Mean living space of the district

Ylemmän%	float64	0.16	Proportion of residents with master's degree to all over 18 years in the area
Naiset,%	float64	0.52	Proportion of women to total residents in the area
kaupunki_Hämeenlinna	uint8	0.03	Hämeenlinna
kaupunki_Joensuu	uint8	0.04	Joensuu
kaupunki_Jyväskylä	uint8	0.08	Jyväskylä
kaupunki_Järvenpää	uint8	0.02	Järvenpää
kaupunki_Kokkola	uint8	0.01	Kokkola
kaupunki_Kouvola	uint8	0.02	Kouvola
kaupunki_Kuopio	uint8	0.07	Kuopio
kaupunki_Lahti	uint8	0.08	Lahti
kaupunki_Lappeenranta	uint8	0.02	Lappeenranta
kaupunki_Mikkeli	uint8	0.01	Mikkeli
kaupunki_Nurmijärvi	uint8	0.02	Nurmijärvi
kaupunki_Oulu	uint8	0.09	Oulu
kaupunki_Pietarsaari	uint8	0.01	Pietarsaari
kaupunki_Pori	uint8	0.03	Pori
kaupunki_Savonlinna	uint8	0.01	Savonlinna
kaupunki_Seinäjoki	uint8	0.02	Seinäjoki
kaupunki_Tampere	uint8	0.13	Tampere
kaupunki_Turku	uint8	0.1	Turku

Table 9 - The variables of the whole dataset after pre-processing

Variable	Type	Mean	Description
Sqm (logarithm)	float64	4.14	Square meters
Price (logarithm)	float64	12.06	Price
rt	int64	0.23	Row house
vuokra	int64	0.28	Rental plot
hyvä	int64	0.67	Good condition
2 h	float64	0.38	2 rooms
3 h	float64	0.3	3 rooms
4 h	float64	0.17	4 rooms
parveke	float64	0.14	Has a balcony
Floor	int64	2.62	Floor level
Year_Very old	int64	0.31	Building year between 1960 - 1979
Year_old	int64	0.23	Building year between 1980 - 1999
Year_fresh	int64	0.23	Building year between 2000 - 2018
Year_new	int64	0.14	Building year 2019 -
Asukkaiden keski-ikä, 2020 (HE)	float64	40.64	Average age of residents
Asukkaiden ostovoimakertymä, 2020 (HR)	float64	227799616.62	Cumulative purchasing power of residents
Pientaloasunnot, %	float64	0.31	Proportion of one-family houses to total dwellings in the area
Ylemmän%	float64	0.17	Proportion of residents with master's degree to all over 18 years in the area
kaupunki_Jyväskylä	uint8	0.1	Jyväskylä
kaupunki_Kuopio	uint8	0.09	Kuopio
kaupunki_Lahti	uint8	0.11	Lahti
kaupunki_Oulu	uint8	0.12	Oulu

kaupunki_Tampere	uint8	0.18	Tampere
kaupunki_Turku	uint8	0.13	Turku

Table 10 - The variables of the large cities' dataset after pre-processing

Variable	Type	Mean	Description
Sqm (logarithm)	float64	4.2	Square meters
Price (logarithm)	float64	11.66	Price
rt	int64	0.31	Row house
vuokra	int64	0.16	Rental plot
hyvä	int64	0.66	Good condition
good_e	int64	0.23	Good energy class
2 h	float64	0.38	2 rooms
3 h	float64	0.31	3 rooms
4 h	float64	0.17	4 rooms
Floor	int64	1.92	Floor level
Year_Very old	int64	0.33	Building year between 1960 - 1979
Year_old	int64	0.27	Building year between 1980 - 1999
Year_fresh	int64	0.24	Building year between 2000 - 2018
Year_new	int64	0.07	Building year 2019 -
Kesämökit yhteensä, 2020 (RA)	float64	84.07	Total number of cottages in the district
Työpaikat yhteensä, 2019 (TP)	float64	3653.17	Total number of jobs in the area
Ylemmän%	float64	0.1	Proportion of residents with master's degree to all over 18 years in the area
kaupunki_Joensuu	uint8	0.17	Joensuu
kaupunki_Järvenpää	uint8	0.1	Järvenpää
kaupunki_Kokkola	uint8	0.05	Kokkola
kaupunki_Kouvola	uint8	0.1	Kouvola
kaupunki_Lappeenranta	uint8	0.09	Lappeenranta
kaupunki_Mikkeli	uint8	0.04	Mikkeli
kaupunki_Nurmijärvi	uint8	0.09	Nurmijärvi
kaupunki_Pietarsaari	uint8	0.02	Pietarsaari
kaupunki_Pori	uint8	0.11	Pori
kaupunki_Savonlinna	uint8	0.03	Savonlinna
kaupunki_Seinäjoki	uint8	0.06	Seinäjoki

Table 11 - The variables of the small cities' dataset after pre-processing

The final pre-processing step was min-max scaling. The independent variables are re-scaled between 0 and 1 to ensure that the variables are in same scale, which reduces algorithms' biased assumptions and enhances the performance.

4.8 Model selection

Objective of this thesis is to find the best performing machine learning models and compare the best performing machine learning models developed with the large and the small Finnish cities datasets. The whole dataset, which contains all observations, and the models developed by using that dataset are used as a benchmark. Next, the model selection process of the three datasets will be presented and the different models will be evaluated. Figure 22 shows the modelling process. After the datasets were cleaned and pre-processed, the datasets were divided into the training and test data. The training dataset contained 80% of the observations and the test data 20%. The training data were used for training the models and the test data were left for the model evaluating purposes. The performance of the models was evaluated by using Mean Absolute Error, Mean Square Error, R-Squared, and Adjusted R-Squared performance metrics which were discussed before.

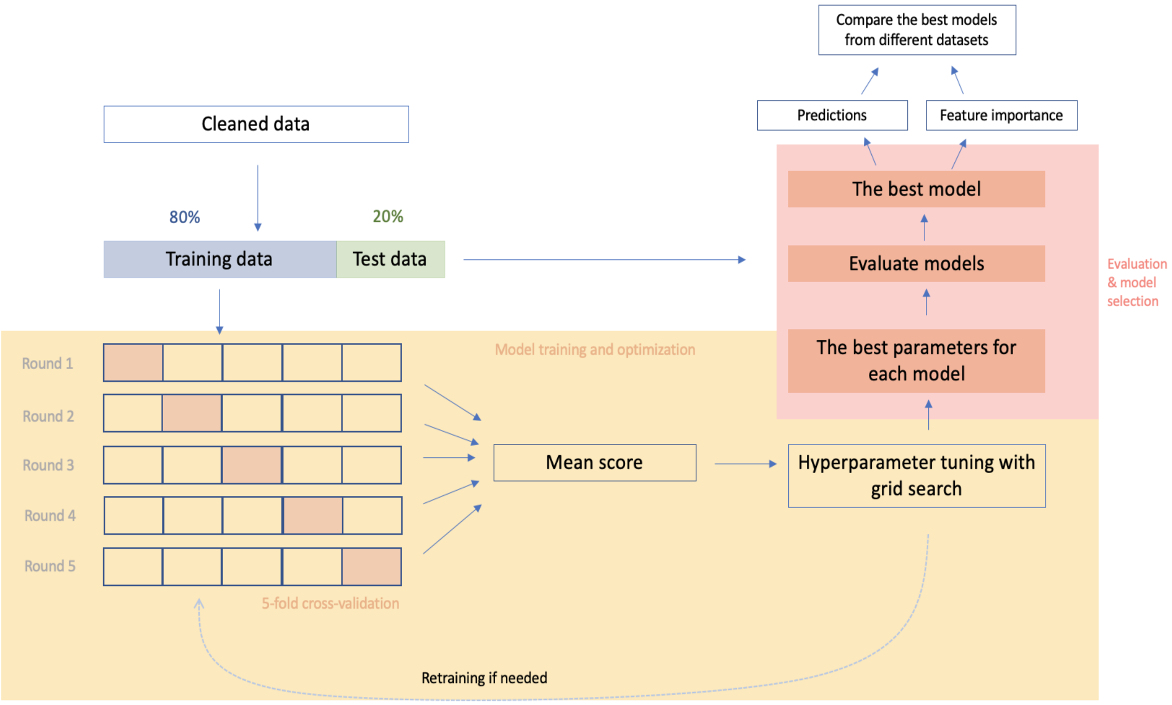

Figure 22 - Modelling process

Tables 12 and 13 present the performance of the models developed by using the whole dataset.

Model name	MAE	R2	Adjusted R2	MSE
XGBoost	15,394.29	0.965	0.965	5.106310e+08
RandomForest	13,215.76	0.970	0.970	4.343144e+08
ExtraTreesRegressor	9,530.54	0.984	0.984	2.252708e+08
DecisionTreeRegressor	11,171.45	0.978	0.978	3.125443e+08
LinearRegression	32,798.75	0.849	0.849	2.172553e+09

Table 12 - Model evaluation with the training data – the whole dataset

Model name	MAE	R2	Adjusted R2	MSE
XGBoost	22,901.70	0.919	0.919	1.175177e+09
RandomForest	25,859.77	0.899	0.899	1.476242e+09
ExtraTreesRegressor	25,314.02	0.900	0.900	1.454303e+09
DecisionTreeRegressor	32,981.62	0.826	0.826	2.533505e+09
LinearRegression	31,760.28	0.869	0.869	1.907228e+09

Table 13 - Model evaluation with the test data - the whole dataset

The LinearRegression model was used as a benchmark. As seen from Table 13, the LinearRegression model has the lowest performance. The other models slightly suffered from overfitting since the performance is better with the training set than with the test set. Nevertheless, the performance of XGBoost, RandomForest, and ExtraTreesRegressor were good also with the test set. The XGBoost model outperformed all the other tree-based models with a good margin. The MAE of the model suggests that model's price predictions have the average error of EUR 22.901. According to R-Squared and Adjusted R-Squared, the model explains 91.9% of the variance of the dependent variable, which is a good result. The best results were achieved by using following parameters:

- booster: gbtree
- learning rate: 0.08
- max depth: 6

- max leaves: 2
- n estimators: 400

The performance of the large cities model is discussed next. Tables 14 and 15 show the performance of the models.

Model name	MAE	R2	Adjusted R2	MSE
XGBoost	15,420.14	0.968	0.968	4.965854e+08
RandomForest	13,916.67	0.969	0.969	4.823857e+08
ExtraTreesRegressor	10,373.13	0.983	0.983	2.693190e+08
DecisionTreeRegressor	27,456.03	0.893	0.893	1.680905e+09
LinearRegression	35,032.25	0.848	0.848	2.393800e+09

Table 14 - Model evaluation with the training data - the large cities

Model name	MAE	R2	Adjusted R2	MSE
XGBoost	24,991.38	0.908	0.908	1.386290e+09
RandomForest	28,295.80	0.876	0.876	1.877623e+09
ExtraTreesRegressor	28,081.92	0.878	0.878	1.856031e+09
DecisionTreeRegressor	36,377.03	0.807	0.807	2.929742e+09
LinearRegression	35,016.80	0.836	0.836	2.491019e+09

Table 15 - Model evaluation with the test data – the large cities

The performance metrics show that the benchmark model, the LinearRegression, underperforms with this dataset, too. Again, the XGBoost model is the best performing one. The explanatory power is weaker than with the whole dataset, for instance R-Squared is lower and MAE is higher. One reason for the higher MAE can be the fact that usually dwellings in larger cities are more expensive. The best results were achieved by using following parameters:

- booster: gbtree
- learning rate: 0.08
- max depth: 6
- max leaves: 2
- n estimators: 400

The hyperparameters were exactly the same than with the whole dataset. The overall performance was weaker than the performance of the whole dataset model.

Tables 16 and 17 show the performance of the models tested by using the small cities dataset.

Model name	MAE	R2	Adjusted R2	MSE
XGBoost	15,569.46	0.932	0.932	4.978875e+08
RandomForest	11,832.19	0.955	0.955	3.273783e+08
ExtraTreesRegressor	13,565.78	0.942	0.942	4.249040e+08
DecisionTreeRegressor	21,062.31	0.873	0.872	9.293266e+08
LinearRegression	26,360.53	0.813	0.813	1.362226e+09

Table 16 - Model evaluation with the training data – the small cities

Model name	MAE	R2	Adjusted R2	MSE
XGBoost	20,030.00	0.889	0.888	8.974838e+08
RandomForest	22,068,94	0.869	0.868	1.057130e+09
ExtraTreesRegressor	23,907.63	0.841	0.840	1.284060e+09
DecisionTreeRegressor	29,372.67	0.773	0.772	1.834731e+09
LinearRegression	25,076.22	0.842	0.842	1.273269e+09

Table 17 - Model evaluation with the test data – the small cities

The overall performance was weaker than the performance of the models discussed earlier. The XGBoost model is the best performing model with this dataset, too. The MAE with this dataset is the lowest of all, which supports the previous conclusion that the prices in larger cities are often more expensive. R-Squared and Adjusted R-Squared metrics are the weakest with this dataset. The obvious reason is that this dataset has fewer observations than the previous ones. We can see that the performance is at high level with the training data, and it drops with the test data. Nevertheless, the performance of the models is still on a good level. The best results were achieved by using following parameters, which slightly differ from prior models:

- booster: gbtree
- learning rate: 0.04
- max depth: 5
- max leaves: 2
- n estimators: 400

The findings in this Chapter demonstrate that the XGBoost algorithm is powerful and suitable for predicting the dependent variable. It consistently outperforms the other algorithms across all datasets. That conclusion answers to the second research question, *'Which of the developed machine learning models is the most reliable in predicting the price of a dwelling in a specific area?'*

4.9 Results

This chapter explains and discusses the results of the study. The performance of the best performing models is evaluated and compared in detail. First, the performance of the best performing whole cities model is examined, following the best performing large cities model, and the best performing small cities model.

4.9.1 The Whole Cities' Model

XGBoost yielded the results shown in Table 18 by using dataset that contained observations from all cities studied in this paper. The error metrices used in the thesis are the same that discussed in Chapter 3. The metrices were calculated in Python by using Scikit learn's sklearn.metrics module which provides Mean Squared Error, Mean Absolute Error, and R-Squared functions. R-Squared was used to create Adjusted R-Squared with the formula presented in Chapter 3.8.3.2.

Model name	MAE	R2	Adjusted R2	MSE
XGBoost	22,901.704	0.919	0.919	1.175177e+09

Table 18 - The whole cities model

Table 19 displays the test set sizes and the corresponding performance metrics on the city level, while Table 23 presents the performance metrics on the square meter bins level. The evaluation on the city level aims to assess the models' performance in individual cities. Additionally, the influence of dwelling size on performance was investigated. Determining the models' performance in this way helps to gain a deeper understanding of the models' effectiveness and facilitates the ability to address the research questions 2 and 4. The observations from individual cities were extracted from the test set and the model was tested by using observations of different cities separately to get the results in Table 19 and Figure 23. The same method was used with other models that are discussed later.

City	MAE	MSE	R-Squared	Test set size
Hämeenlinna	15,523.57	4.523262e+08	0.889	63
Joensuu	17,647.14	5.331040+08	0.807	102
Jyväskylä	21,091.81	8.492467e+08	0.863	184
Järvenpää	19,562.73	6.736579e+08	0.871	58
Kokkola	19,165.12	5.689588e+08	0.863	26
Kouvola	16,122.06	5.166146e+08	0.833	46
Kuopio	22,844.77	1.208825e+08	0.801	139
Lahti	20,510.26	1.525779e+09	0.832	182
Lappeenranta	17,370.35	6.512607e+08	0.808	45
Mikkeli	19,809.49	5.981971e+08	0.808	31
Nurmijärvi	33,291.02	2.379882e+09	0.841	64
Oulu	21,531.09	8.929759e+08	0.855	195
Pietarsaari	9,568.89	1.142866e+08	0.982	8
Pori	18,327.63	6.061186e+08	0.848	71
Savonlinna	13,739.97	6.061186e+08	0.550	19
Seinäjoki	19,690.97	5.798164e+08	0.716	39
Tampere	20,571.07	9.634768e+08	0.872	282

Turku	22,868,78	9.579501e+08	0.875	185
Espoo	32,718.24	2.152210e+09	0.872	440

Table 19 - Test set results of the whole cities model – city level

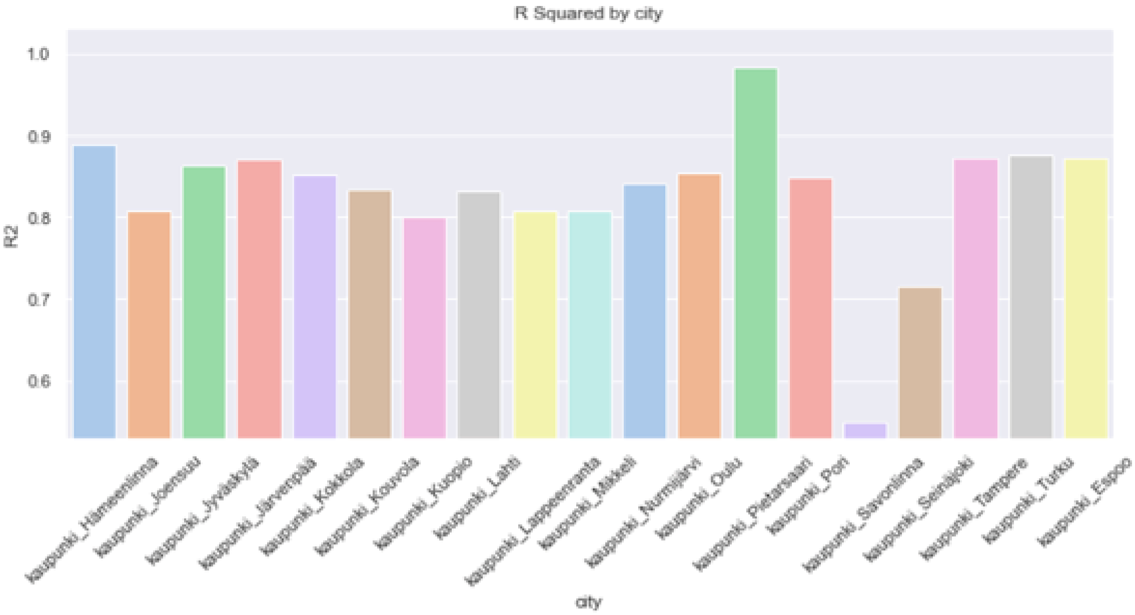

Figure 23 - R-Squared by city - The whole cities dataset

Figure 23 shows that the model explained the variance fairly well on a city level. However, the city of Savonlinna was difficult to predict since the model achieved R-Squared below 60% with the city. The highest R-Squared is 0.982 and the lowest MAE is € 9,569 for dwellings in Pietarsaari.

One reason for the poor fit of Savonlinna and the good fit of Pietarsaari might be the small test data set size in these cities. Even though, MAE and MSE of Savonlinna are good. On the other cities, the model performs well, albeit there is performance variance between the cities.

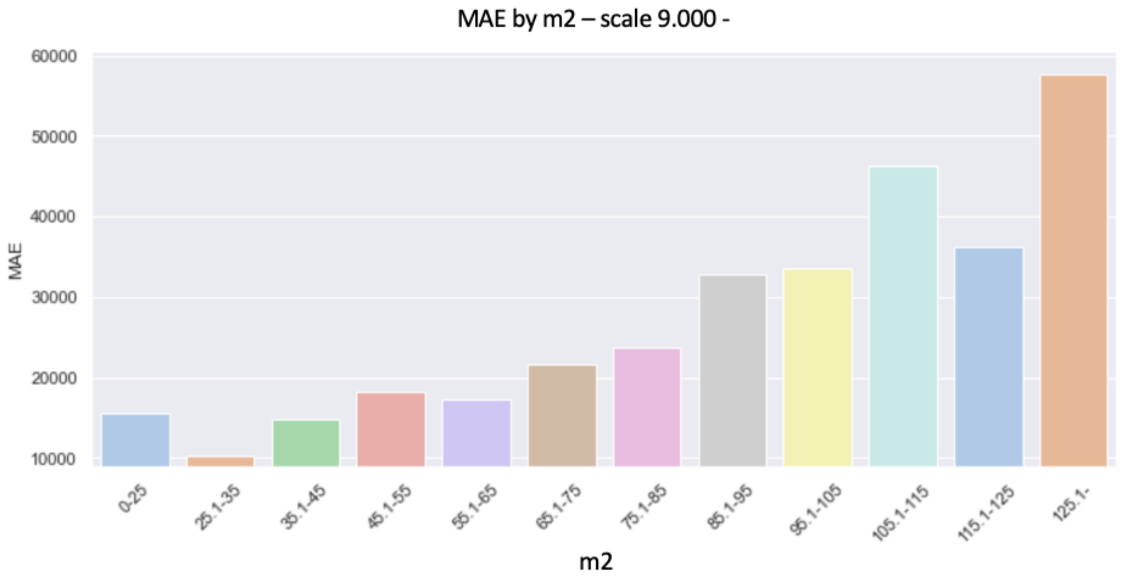

Figure 24 - MAE by square meters - The whole cities dataset

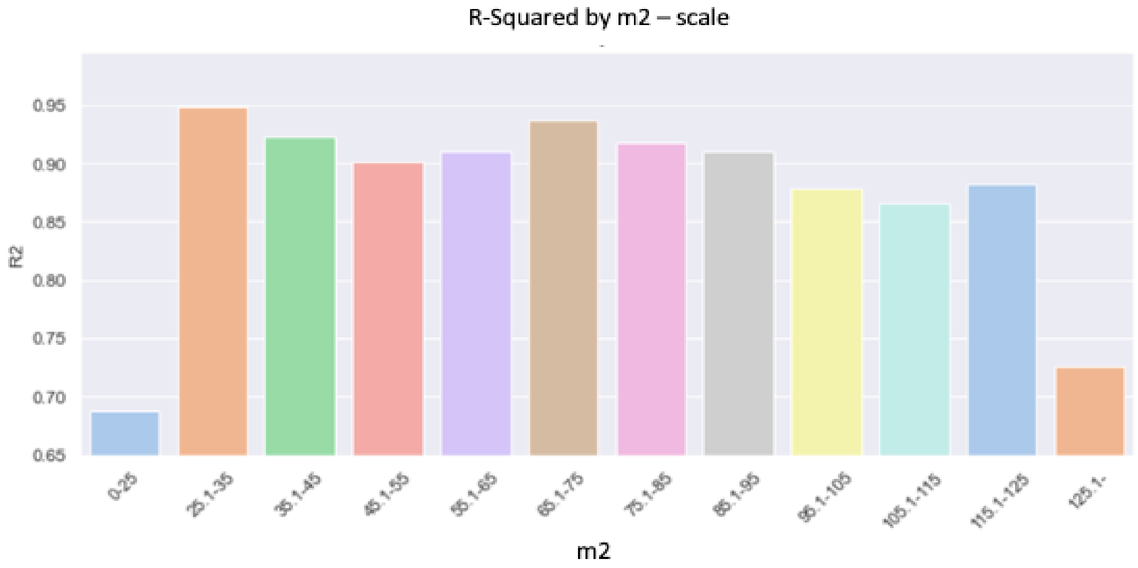

Figure 25 - R-Squared by square meters - The whole cities dataset

Figures 24, 25 and Table 20 show that the size of a dwelling impacts on the model's performance. The best R-Squared is 0.948, MAE is € 10,332, and MSE is 1.977430e+08 which were achieved in the 25.1 to 35 square meter interval. The worst R-Squared is 0.688 being in the 0 - 25 square meter interval. The highest MSE is € 57,645 and MSE 6.866357e+09, both achieved in the 125.1 – square meter interval, which presents the largest dwellings of this study. Typically, the larger dwellings are more expensive, which affects the largest dwellings' MAE and MSE results. Despite that, R-Squared of the largest dwellings is the second worst 0.726. The model predicts well the prices of the dwellings which square meters are between 25 – 125, while the predicting power is weaker with the smallest and the largest dwellings.

Square meters	MAE	MSE	R-Squared	Test set size
0 - 25	15,587.80	6.288751e+08	0.688	20
25.1 - 35	10,332.13	1.977430e+08	0.948	213
35.1 – 45	14,798.99	3.660588e+08	0.924	231
45.1 - 55	18,273.27	5.912786e+08	0.901	278
55.1 - 65	17,299.92	5.645205e+08	0.911	404
65.1 - 75	21,612.65	7.691658e+08	0.937	262
75.1 - 85	23,707.86	1.012719e+09	0.918	254
85.1 - 95	32,894.92	1.811283e+09	0.911	158
95.1 - 105	33,633.91	1.986630e+09	0.879	121
105.1 - 115	46,372.27	3.442592e+09	0.866	67
115.1 - 125	36,211.91	2.064411e+09	0.883	66
125.1 -	57,645.07	6.866357e+09	0.726	105

Table 20 - Test set results of the whole cities model – square meters

The third research question is *'Which features of a dataset are the most significant in predicting the price of a dwelling?'* In order to answer the question, the importance of the independent variables was calculated by using XGBoost's feature_importances_ function. The feature importance metric quantifies the extent to which each independent variable contributed to the construction of the boosted decision trees within the model. The measure reflects the degree of significance that a given variable holds in relation to the others. Furthermore, the relative importance of an independent variable amplifies as it has a more pronounced role in key decisions. First, the feature importance is calculated for every attribute in a single decision tree. After that, the mean scores of the independent variables' feature importance in the decision trees within the model are extracted (Brownlee 2016).

The default calculation method, which is *'gain'* method, was used to calculate the feature importance. The gain metric represents the mean gain that ensues from every partition in which a feature was employed (Płoński 2020). Higher the score, which is represented on the x-axis in Figure 26, more important the feature is in the model.

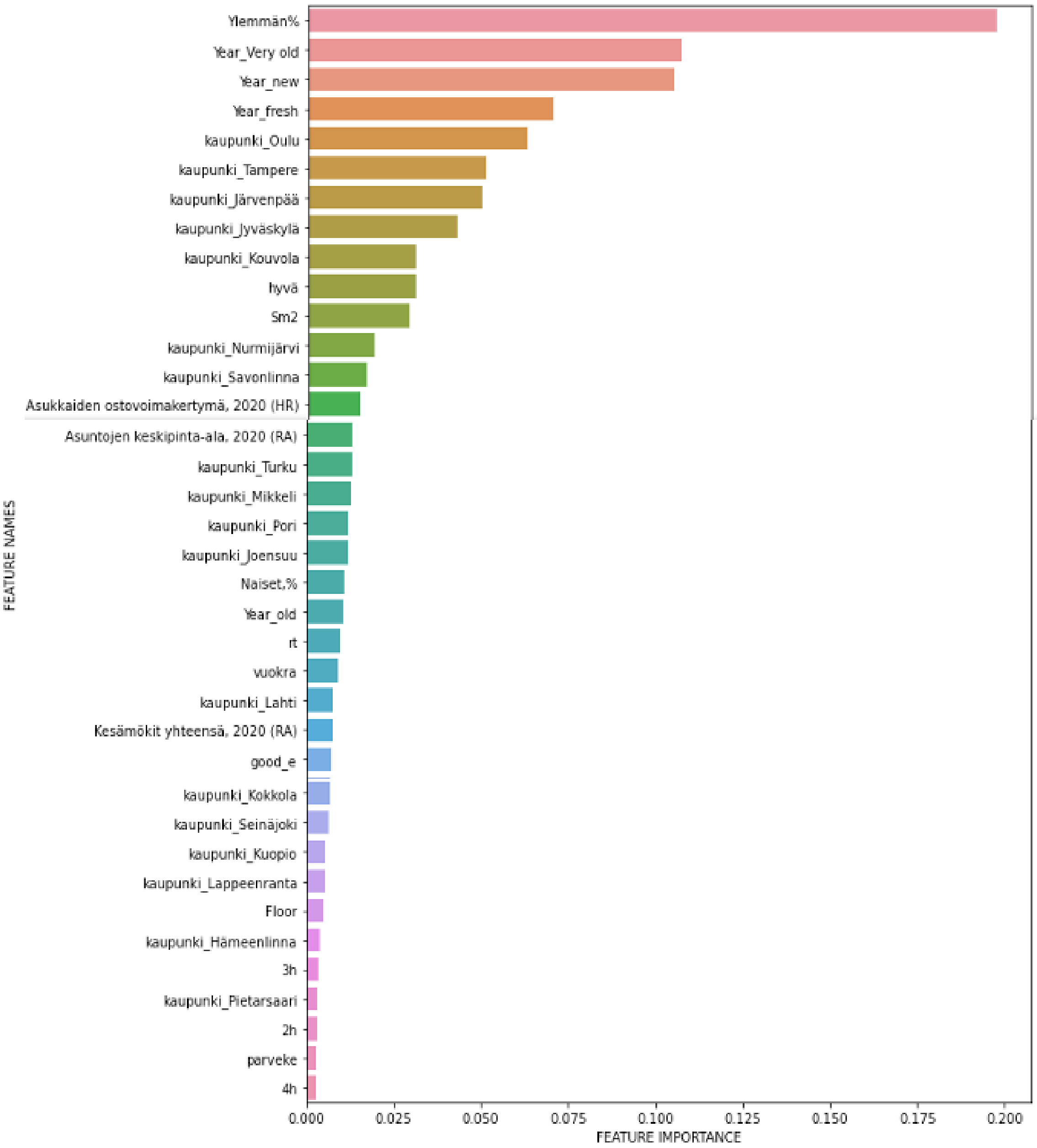

Figure 26 - The feature importance of the whole cities model

The most important variable for the dwelling price prediction of this model is the Ylemmän% variable that explains the proportion of residents having a master's degree in a district. After that, the next important variables are building year variables. The location variables that describe the city of the observation are the next important. However, some of the cities have more importance on the model than others. The square meters variable is only the 11th important, but still having importance. In addition, the good condition variable *'hyvä'* has a moderate contribution. The number of rooms, balcony, and floor variables seem to have a low importance. As discussed in Chapter 2, spatial features impact on dwelling prices and in this case a spatial feature seems to have the highest importance.

4.9.2 The Large Cities' Model

The results achieved with the large cities model are discussed next.

City	MAE	MSE	R-Squared	Test set size
Jyväskylä	24,863.01	1.400024e+09	0.803	149
Kuopio	23,189.61	1.190951e+09	0.836	129
Lahti	23,852.57	1.353734e+09	0.846	193
Oulu	22,343.28	1.019572e+09	0.825	218
Tampere	22,397.55	1.257137e+09	0.8402	285
Turku	24,834.49	1.301297e+09	0.827	204
Espoo	30,465.21	1.886045e+09	0.883	449

Table 21 - Test set results of the large cities model – city level

The large cities model achieved good results, too. The model explains circa 91% of the variance in the dependent variable. However, the overall performance is slightly weaker than the overall performance of the whole cities model, which is probably due to the smaller training set size. XGBoost achieved the following results presented in Table 22. The test set sizes on a city level are presented in Table 21.

Model name	MAE	R2	Adjusted R2	MSE
XGBoost	24,991.382	0.908	0.908	1.386290e+09

Table 22 - The large cities model

Figures 27 and 28 present the model's performance on a city level. The performance was particularly good in every city. R-Squared and MAE were highest in Espoo. Espoo had most observations in the training data and the highest mean price, as we see from Table 1, hence the model might have influenced too much of the observations from Espoo. The performance of this model is better only in Espoo and Lahti compared to the whole cities model. The best R-Squared is 0.882 achieved in Espoo. The lowest MAE and MSE are € 22.343 and 1.019572e+09, respectively, for dwellings in Oulu. MAE is low in all cities, for instance, Espoo is the only city where MAE is above € 30,000. The results recommend that the model is suitable for the dwelling price prediction in these cities, even though it might be slightly biased.

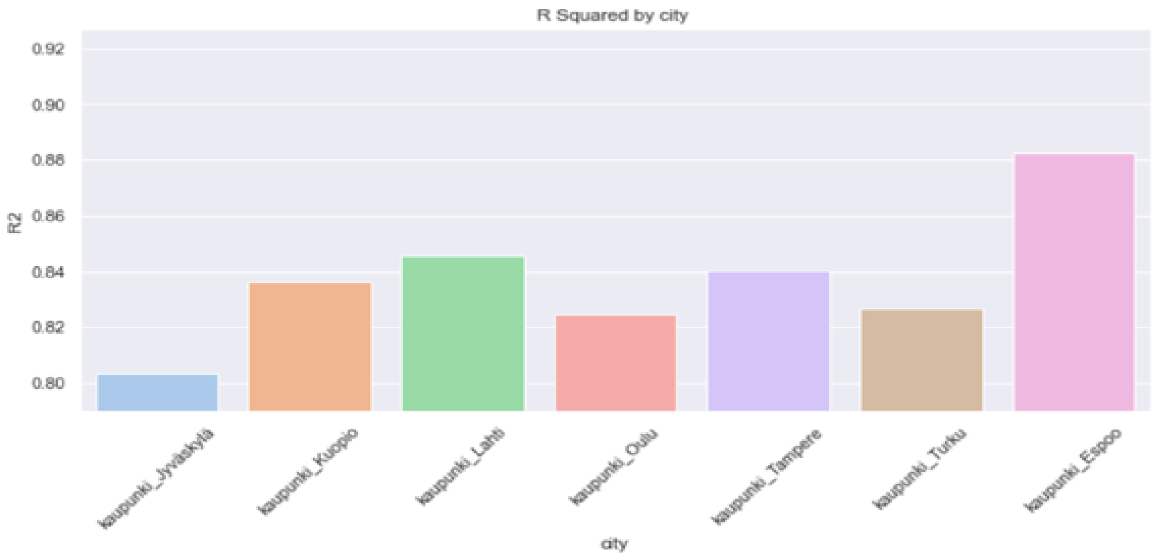

Figure 27 - R-Squared by city - The large cities model

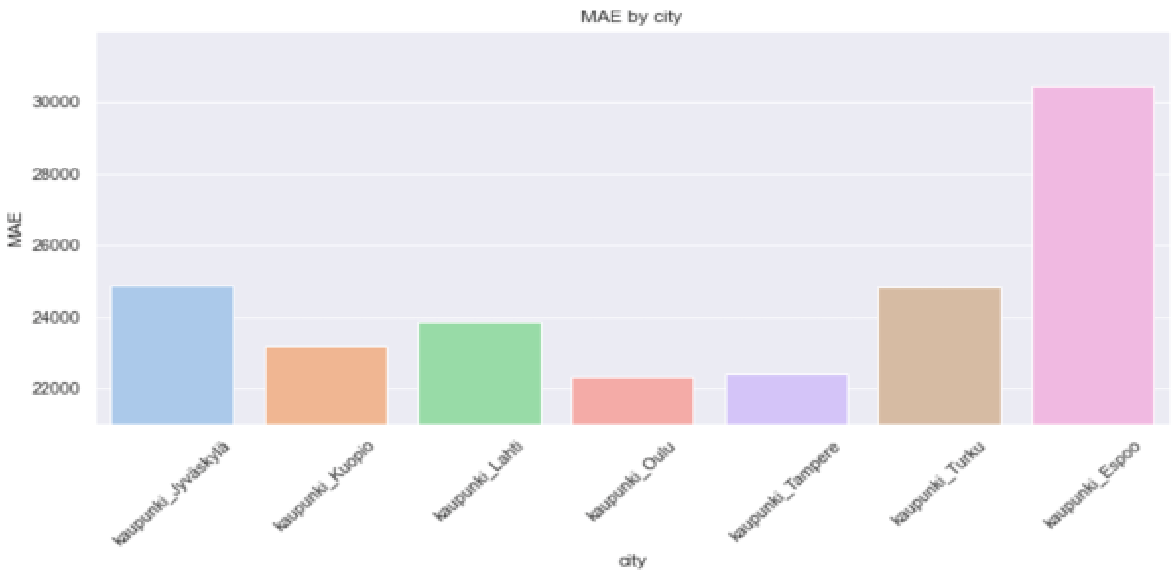

Figure 28 - MAE by city - The large cities model

As seen in Figures 29 and 30 and in Table 23, the performance of the large cities model follows the same kind of pattern than the performance of the whole cities model. The model works better with smaller dwellings and the performance starts to decrease with larger dwellings. The model's performance with the dwellings between 0-25 square meters is better than the performance of the whole dataset model and the performance of the model is at sufficient level when the dwelling size is between 25.1 to 125 square meters. The performance with dwellings over 125 square meters is weak.

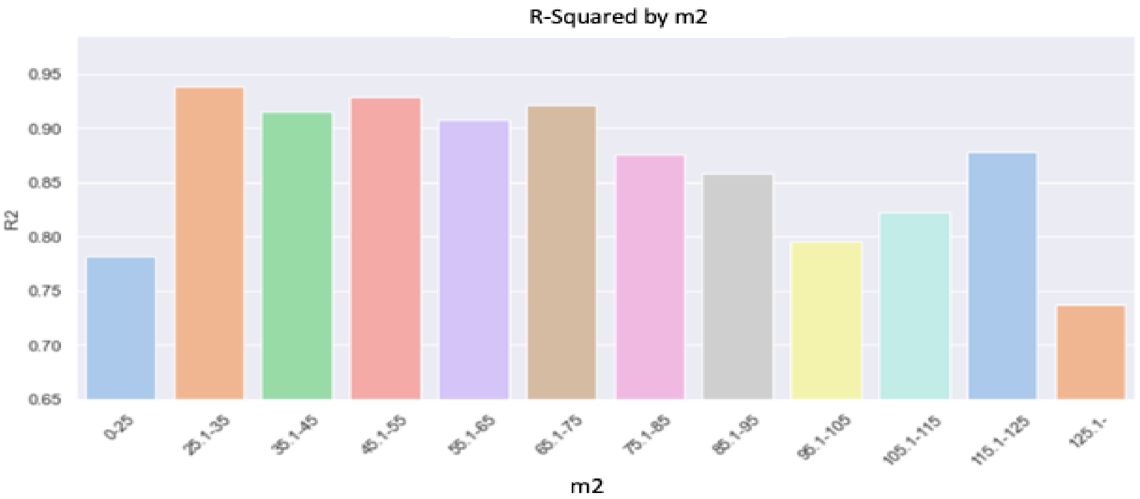

Figure 29 - R-Squared by square meters - The large cities model

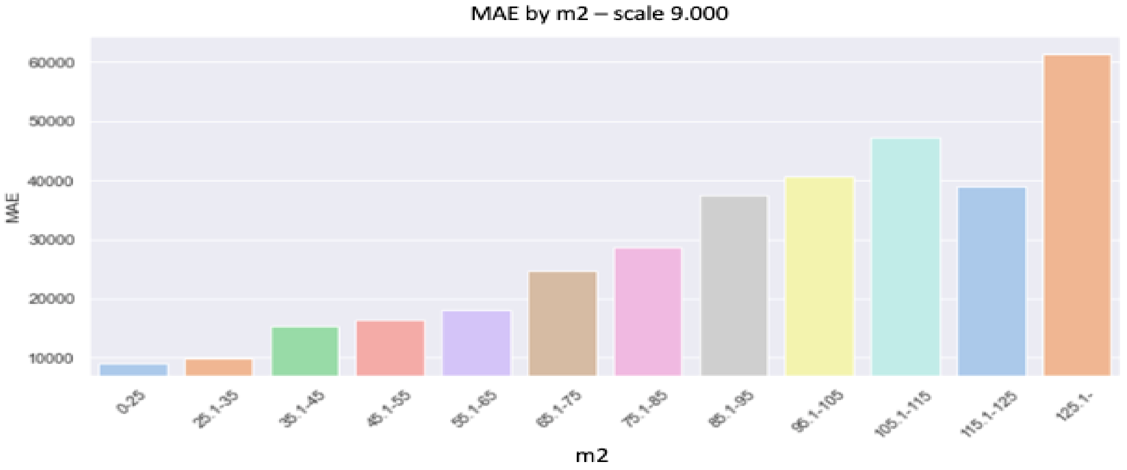

Figure 30 - MAE by square meters - The large cities model

Square meters	MAE	MSE	R-Squared	Test set size
0 - 25	9,067.12	1.698980e+08	0.782	21
25.1 - 35	9,813.49	1.735162e+08	0.939	156
35.1 – 45	15,306.89	4.654572e+08	0.916	150
45.1 - 55	16,473.46	4.317597e+08	0.928	198
55.1 - 65	18,154.38	5.895296e+08	0.909	286
65.1 - 75	24,610.45	1.057787e+09	0.922	212
75.1 - 85	28,729.02	1.625639e+09	0.876	214
85.1 - 95	37,432.99	2.277203e+09	0.859	139
95.1 - 105	40,599.79	3.013724e+09	0.795	74
105.1 - 115	47,329.60	3.467938e+09	0.823	34
115.1 - 125	38,925.80	2.408668e+09	0.878	50
125.1 -	61,280.71	6.578442e+09	0.737	93

Table 23 - Test set results of the large cities model – square meter bins

Figure 31 explains the feature importance of the large cities model. The feature importance is similar than in the whole cities model. The Ylemmän% is the most important feature following building year features. After them, are the location-based variables. The square meters and the condition of a dwelling have importance. The number of rooms, balcony, and floor variables have a low importance, similarly as we have seen with the whole cities model. Spatial features seem to have high importance on large cities of Finland.

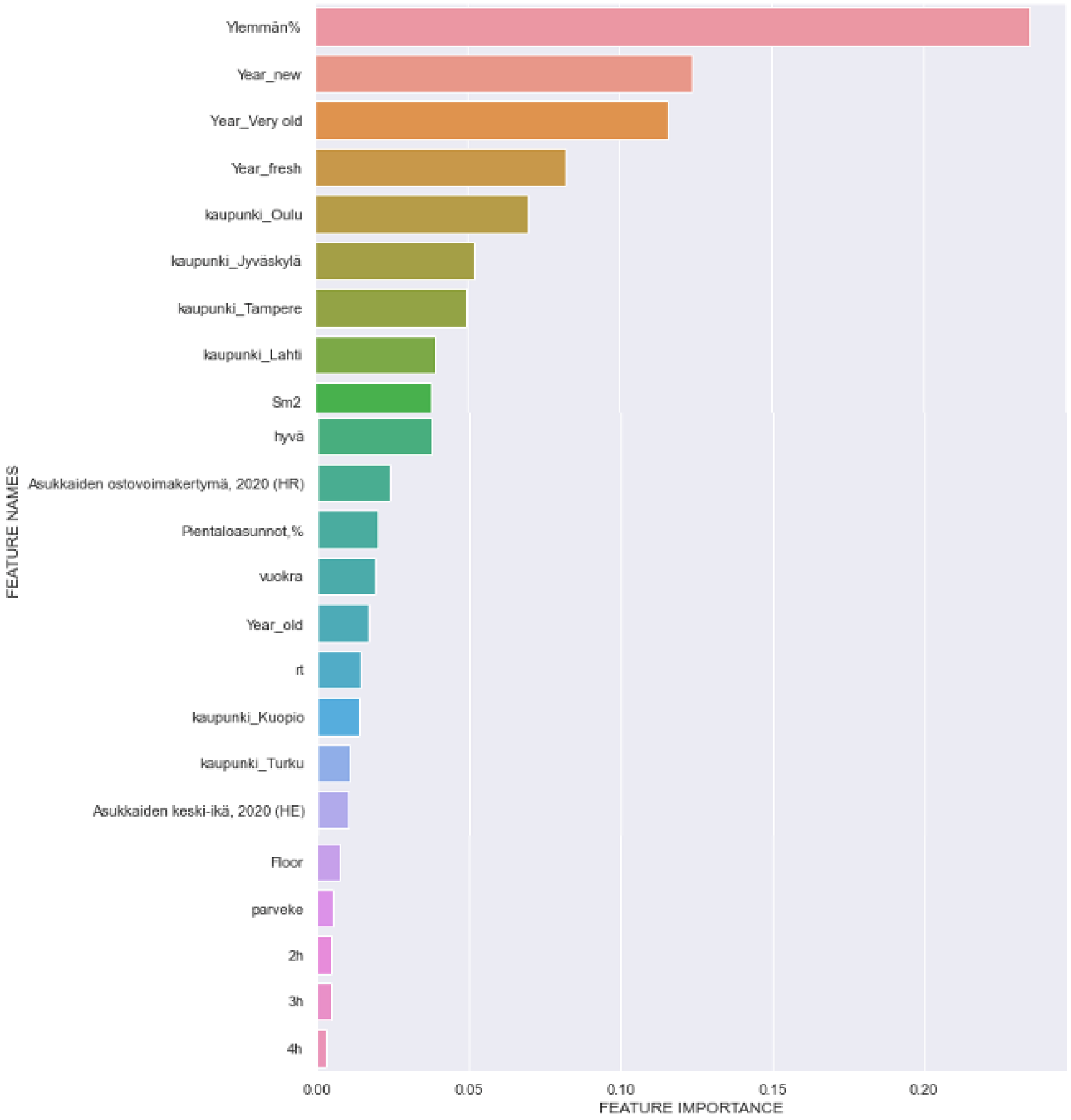

Figure 31 - The feature importance of the large cities model

4.9.3 The Small Cities' Model

Last, the performance of the small cities model is discussed. The small cities model had the lowest performance. Most obvious reason is the smallest training set size. XGBoost achieved the following results presented in Table 24. The test set sizes and the performance metrices on a city level are presented in Table 25.

Model name	MAE	R2	Adjusted R2	MSE
XGBoost	20,030.000	0.889	0.888	8.974838e+08

Table 24 - The performance of the small cities model

City	MAE	MSE	R-Squared	Test set size
Hämeenlinna	20,797.42	7.761890e+08	0.886	66
Joensuu	22,460.86	1.278774e+09	0.730	85
Järvenpää	21,248.01	9.860088e+08	0.862	58
Kokkola	16,726.87	5.343453e+08	0.849	27
Kouvola	18,324.64	5.829352e+08	0.698	59
Lappeenranta	21,381.22	1.239706e+09	0.862	45
Mikkeli	16,126.95	4.254130e+08	0.911	20
Nurmijärvi	29,273.05	1.611582e+09	0.882	47
Pietarsaari	23,815.55	8.762855e+08	0.805	20
Pori	18,529.93	6.720425e+08	0.837	72
Savonlinna	12,981.82	2.827737e+08	0.597	22
Seinäjoki	13,968.49	3.626602e+08	0.857	31

Table 25 - Test set results of the small cities model – city level

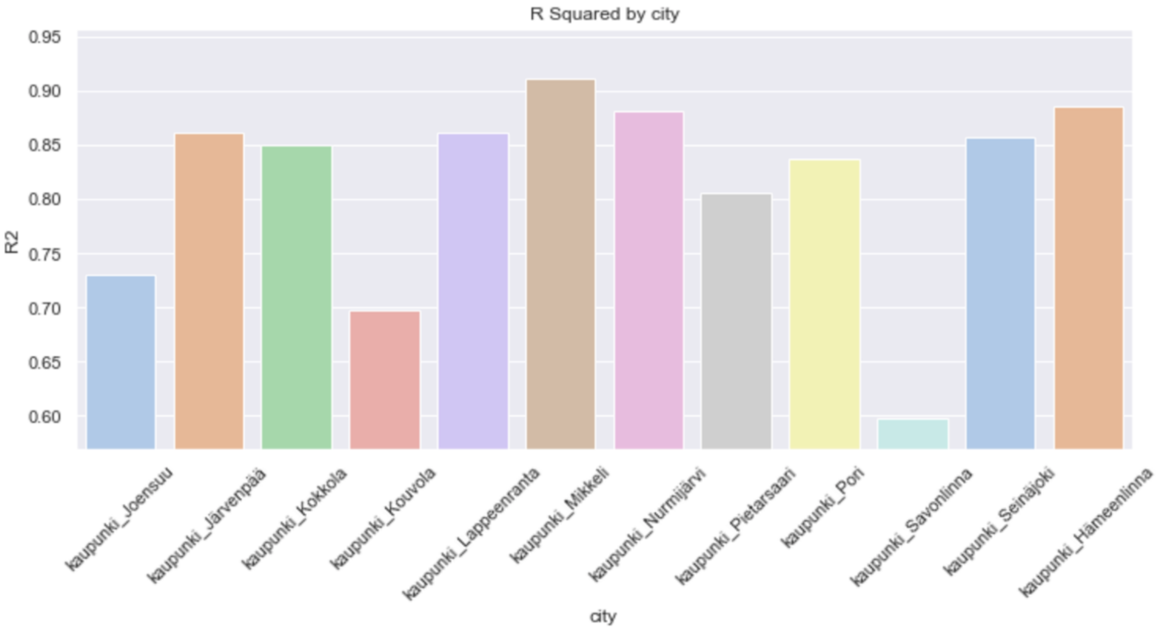

Figure 32 - R-Squared by city - The small cities model

Figures 32 and 33 show that there is variance in terms of the performance of the small cities model on a city-level. R-Squared of Järvenpää, Kokkola, Lappeenranta, Mikkeli, Nurmijärvi, Seinäjoki, and Hämeenlinna are at good levels while R-Squared of some other cities, e.g., Kouvola and Savonlinna are significantly weaker. Mean Absolute Errors are at lower levels, except MAE of Nurmijärvi, than MAEs of the large cities model. Lower dwelling price levels in smaller cities can explain that phenomenon. The errors given in euros are low, which were the aim. The model performs better in Mikkeli, Nurmijärvi, Savonlinna, and Seinäjoki than the whole cities model, while in the other cities the performance is weaker.

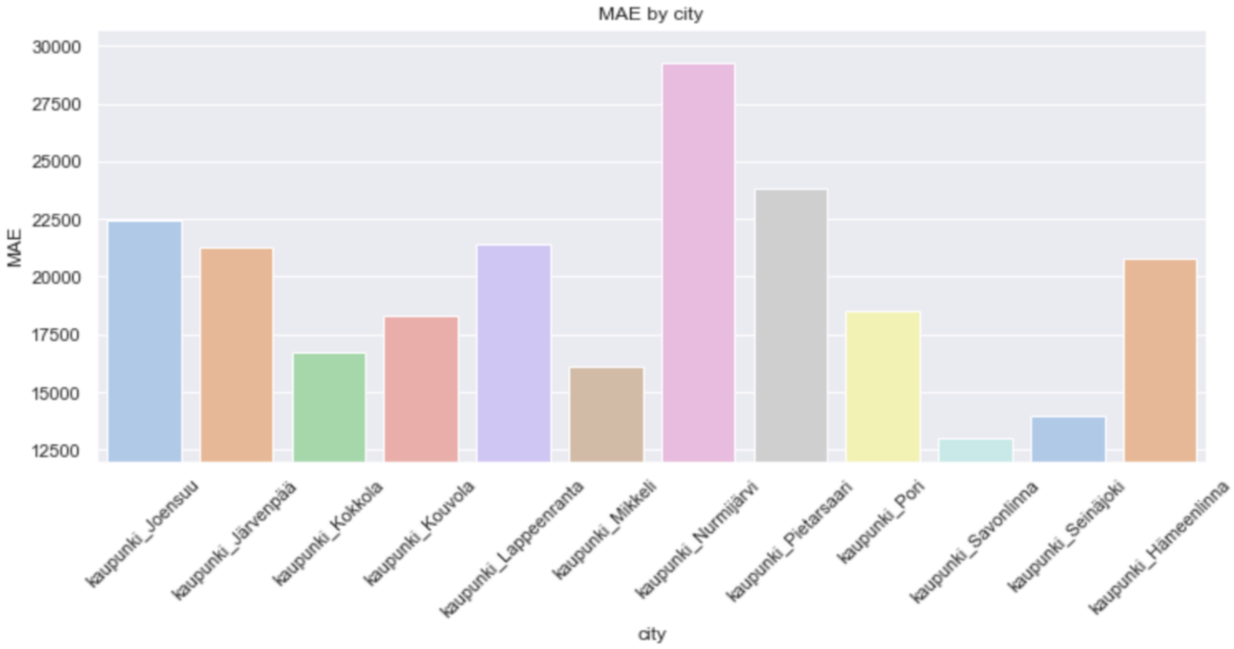

Figure 33 - MAE by city - The small cities model

Square meters	MAE	MSE	R-Squared	Test set size
0 - 25	5,536.35	5.611922e+07	0.944	5
25.1 - 35	10,225.51	1.909117e+08	0.832	45
35.1 – 45	11,385.25	2.146173e+08	0.896	35
45.1 - 55	15,031.03	3.308687e+08	0.875	56
55.1 - 65	15,689.67	4.258112e+08	0.892	114
65.1 - 75	17,962.86	7.518741e+08	0.874	68
75.1 - 85	19,295.57	5.789555e+08	0.901	77
85.1 - 95	25,478.71	9.459530e+08	0.796	39
95.1 - 105	27,401.71	1.207101e+09	0.774	38
105.1 - 115	29,152.99	1.199464e+09	0.870	16
115.1 - 125	30,527.86	2.282963e+09	0.757	22
125.1 -	49,715.25	4.270512e+09	0.721	37

Table 26 - Test set results of the small cities model – square meter bins

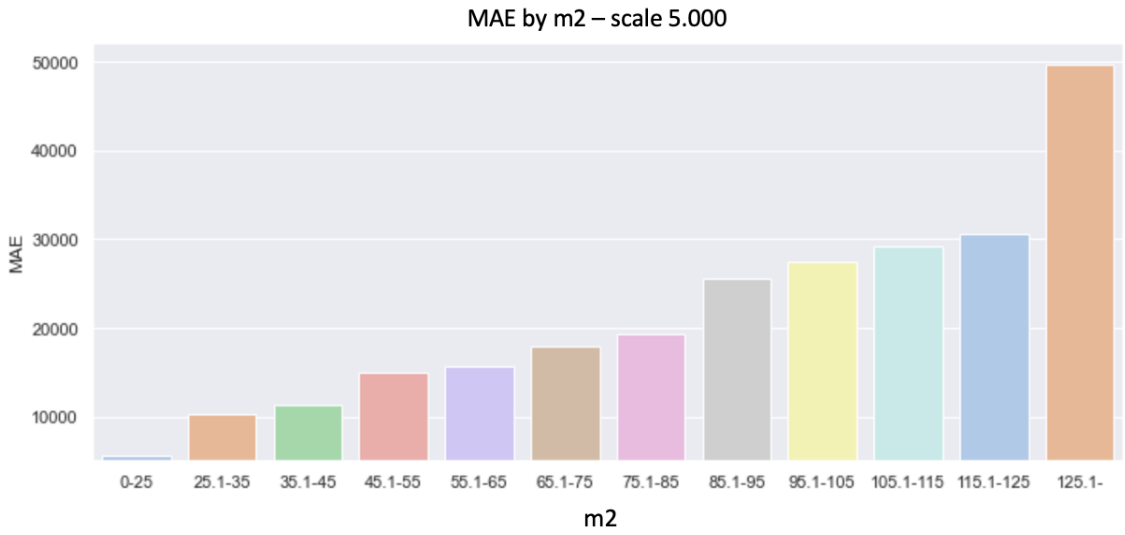

Figure 34 - MAE by square meters - The small cities model

As seen with the prior models, the small cities model also performs better with smaller dwellings compared to larger dwellings. Again, the model has difficulties to predict the prices of the largest dwellings. The performance of the model seems to be really good with smaller than 25 square meters dwellings. However, the test set of 0-25 square meters dwellings contains only 5 observations which can explain the good performance. The performance is presented in Figures 34 and 35. In overall, the performance is at a good level, since MAE of the smaller than 85 square meters dwellings is below € 20,000 and rise circa € 30,000 with the larger dwellings except the largest.

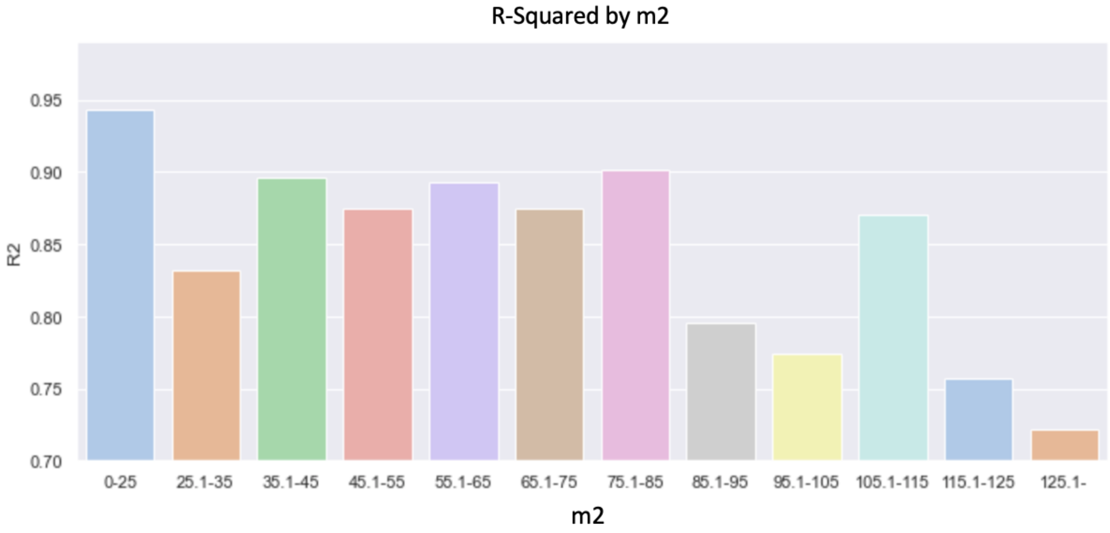

Figure 35 - R-Squared by square meters - The small cities model

As seen from Figure 36, the building year variables are the most important variables of the small cities model. As in the previous models, the number of rooms variables have a low importance. The other low contributing variables in the previous models were the balcony and the floor variables which were removed already in the earlier stages of the small cities model construction due to low impact on the dwelling prices or high correlation with other independent variables. The Ylemmän% variable does not have similar importance on this model that it had on the previous models. Some of the location variables have high importance while the contribution of other location-based variables is lower. The condition and the square meters variables have importance as they had on the other models.

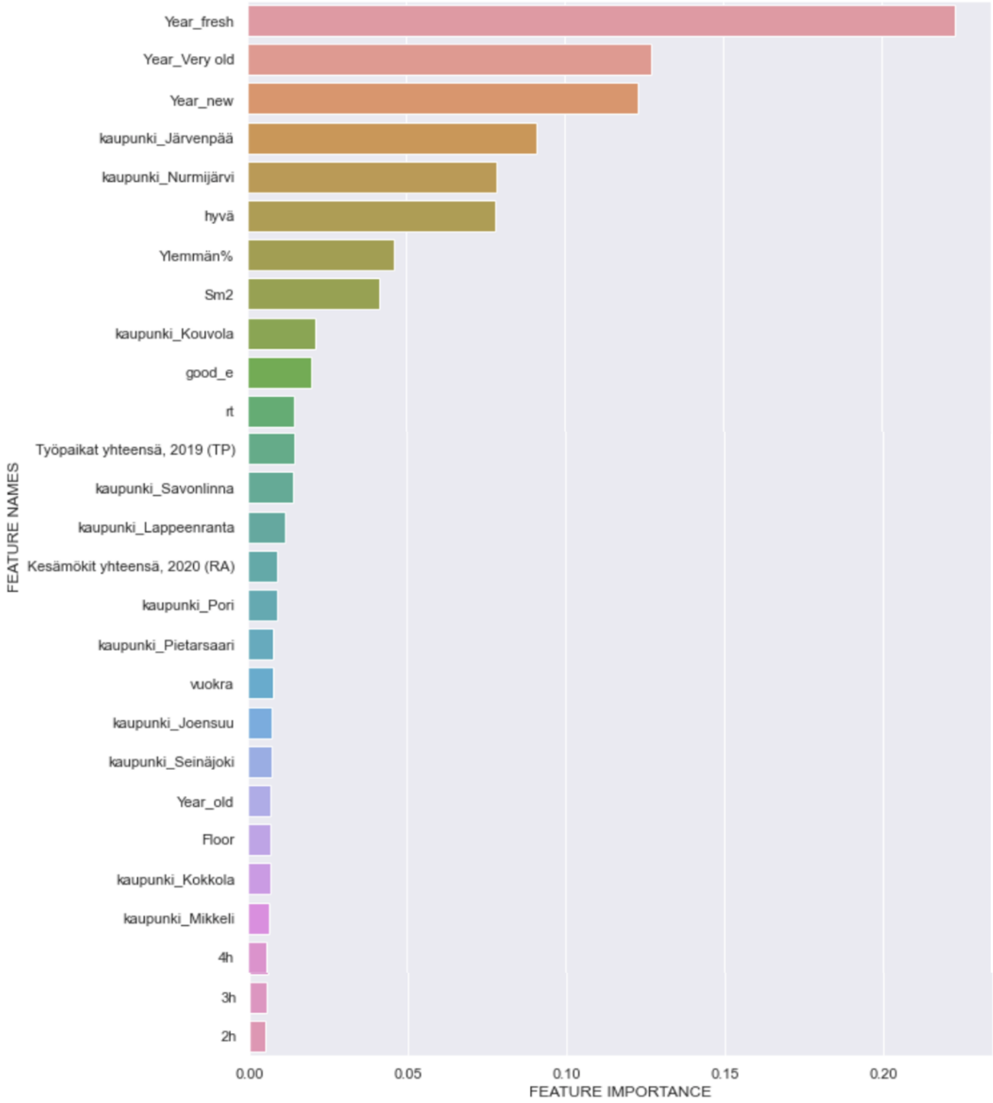

Figure 36 - The feature importance of the small cities model

5 CONCLUSION

The objective of the thesis was to create an understanding of the dwelling markets and investigate the potential of machine learning techniques in predicting the Finnish dwelling prices. Furthermore, the objective was to compare whether the models that are trained with data from different sized Finnish cities have discrepancies. Even though, the Finnish dwelling markets have been studied broadly, the prior studies mostly concentrate on the Helsinki Metropolitan Area. Furthermore, the use of machine learning techniques in the Finnish dwelling market has not been studied comprehensively. Therefore, the study concentrates on multiple Finnish cities with different sizes and characteristics. The study presents the way how a price prediction model for the Finnish dwellings can be constructed by using publicly available data. In addition, it provides evidence on the best performing machine learning algorithms for the price prediction tasks in different sized cities of Finland.

The research questions for the study were:

1. How can machine learning be used to predict the price of a residential building in Finland?
2. Which of the developed machine learning models is the most reliable in predicting the price of a dwelling in a specific area?
3. Which features of a dataset are the most significant in predicting the price of a dwelling?
4. Which are the most significant discrepancies between the models developed in different areas?

Chapters 2 and 3 create a comprehensive understanding of the dwelling markets and machine learning. Chapter 4 presents the way how this knowledge can be used to create actual machine learning models for predicting the Finnish dwelling prices by scraping public data from the Asuntojen.hintatiedot.fi and the Paavo -databases.

Based on the results of the empirical part of this thesis, the answer to the second research question is the XGBoost based models. The algorithm outperforms the other tested algorithms with all test datasets and all evaluation metrics used in this thesis. Some of the algorithms achieved better results with the training data than XGBoost, but with the out-of-sample data XGBoost were superior, achieving 0.919, 0.909, and 0.888 Adjusted R-Squared with the whole, large and small cities datasets, respectively.

The third research question, '*Which features of a dataset are the most significant in predicting the price of a dwelling?'*, can be answered based on the knowledge created in Chapter 4. Figures 26, 31, and 36 present the most important features of every model created. The most important features seem to depend on the size of the city in question. The models indicate that the building year of a dwelling has high importance on the price prediction tasks. On the contrary, the number of rooms, balcony and floor variables have low importance. The obvious reason for the low importance of the number of rooms variables is that the significance of the square meter variable is higher than average, and it has relatively high correlation with the number of rooms variables. The location of the dwelling has importance as Kiel and Zabel (2004) explained. In addition, the condition and square meters variables tend to have importance on the models. Furthermore, spatial variables have importance as we have seen from the contribution of the Ylemmän% variable. This result is in line with the prior studies.

Figures 31 and 36 present the feature importance of the large and the small cities models, respectively. As discussed before, The Ylemmän% independent variable, which explains the proportion of the residents in the area that have a higher academic degree, tends to have more importance on the price prediction in the large Finnish cities than the other variables have. These findings indicate that the proportion of residents with higher academic degree in a district is a significant determinant of the dwelling prices in the large Finnish cities. While the variable retains its importance in smaller cities, its impact on the dwelling prices is relatively weaker compared to that of the other variables. The proportion of jobs and the proportion of cottages in a district variables seem to have importance in small Finnish cities while they do not have importance in large cities. This answers the fourth, and last, research question, '*Which are the most significant discrepancies between the models developed in different areas?'*

Why is the Ylemmän% variable important in the large cities but not so important in the small cities? One reason may be, that the higher proportion of the cities in the large dataset have a university, which might affect the number of residents having a master's degree in a city. Residents with higher education might prefer some neighborhoods which would make the variable Ylemmän% to be an important determinant in the price formation in those cities. Six out of seven cities have a university in the large cities' dataset, while 4 out of 12 has a university campus in the small cities' dataset. However, almost every city in this study have at least a campus of a university of applied sciences if it does not have a university campus.

5.1 Limitations and Future Research

The research in this paper does not cover every city in Finland and the cities chosen to the thesis are decided by the author, which can cause biases to the models presented. In addition, the data used in the thesis is limited. The data contain observations that are collected by humans, which could have led to errors. The large cities dataset contains over 8,000 observations which can be considered sufficient for machine learning models. The small cities' dataset has less than 3,000 observations which is sufficient, too, but the performance is considerably weaker than the performance of the large cities' dataset. The data was collected during circa one-year period. The longer collecting period could lead to a larger dataset, i.e., more comprehensive training sets for the model building. The data were gathered from public data which have limitations; thus, more comprehensive datasets could have been collected by using different sources.

Furthermore, the sales data cover observations from March 2021 to October 2022. That period had several major economical and geopolitical events that have impacted on Finnish dwelling prices, e.g., the COVID-19 pandemic, Russian invasion of Ukraine, Europe's energy crisis, and interest rate hikes due to high inflation. The effects on prices were briefly discussed in Chapter 2, but this thesis does not investigate deeply the effects of those events which could be a potential topic for future research.

The algorithms tested in this thesis contain algorithms that have been proven in earlier studies to be efficient in the dwelling price prediction tasks. However, the algorithms are limited to the algorithms that can be considered as interpretable algorithms due to the nature of the thesis. The author excluded all black-box models, such as deep neural network models from the research, even though, such models can perform well in tasks in question.

As seen from the models developed, the performance of the models on a city-level fluctuates among the models. In order to enhance the model’s performance, the models could be created to every city individually. However, that will require more data and more work.

REFERENCES

Adetunji, A. B. Akande, O. N. Ajala, F. A. Oyewo, A. Akande, Y. F. Oluwadara, G. House Price Prediction using Random Forest Machine Learning Technique. Procedia computer science.

Ali, J. Khan, R. Ahmad, N & Maqsood, I. 2012. Random Forests and Decision Trees. IJCSI International Journal of Computer Science Issues.

Allwright, S. 2022. What is a good MAE score? (simply explained). Retrieved 4.2.2023 from https://stephenallwright.com/good-mae-score/.

Alpaydin, E. 2014. Introduction to machine learning. MIT Press [2014] Third edition.

Alpaydin, E & Bach, F. 2019. Introduction to Machine Learning, MIT Press.

Alruhaymi, A.Z. and Kim, C.J. (2021) Study on the Missing Data Mechanisms and Imputation Methods. Open Journal of Statistics, 11, 477-492. https://doi.org/10.4236/ojs.2021.114030

Anand, S. Yadav, P. Gaur, A. & Kashyap, I. 2021. Real Estate Price Prediction Model. 3rd International Conference on Advances in Computing, Communication Control and Networking.

ArcGIS Pro. n.d. How Extra trees classification and regression algorithm works. Retrieved 8.3.2023 from https://pro.arcgis.com/en/pro-app/latest/tool-reference/geoai/how-extra-tree-classification-and-regression-works.htm

Arrieta, A. B. Díaz-Rodrígues, N. Del Ser, J. Bennetot, A. Tabik, S. Barbado, A. Garcia, S. Gil-Lopez, S. Molina, D. Benjamins, R. Chatila, R. Herrera, F. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Indormation Fusion. Elsevier.

Aurelien, G. 2019. Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow. O'Reilly.

Brounen, D. & Kok, N. 2011. On the economics of energy labels in the housing market. Journal of Environmental Economics and Management. Elsevier.

Brownlee, J. 2016. Feature Importance and Feature Selection with XGBoost in Python. Machine Learning Mastery.

Brownlee, J. 2020. Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python. Machine Learning Mastery.

Cellmer, R. Cichulska, A. & Bełej, M. 2020. Spatial Analysis of Housing Prices and Market Activity with the Geographically Weighted Regression. International Journal of Geo-Information

Cunningham, P. Cord, M & Delany, S. J. 2008. Machine Learning Techniques for Multimedia. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.457.4869&rep=rep1&type=pdf

Chen, T. Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Dai, L. Sheng, X. 2021. The Impact of Uncertainty on State-Level Housing Markets of the United States: The Role of Social Cohesion. Sustainability 2021,13

Deng, Y. Gu, Q. & He, J. 2021. Reinforcement learning and mortgage partial prepayment behavior. Pacific-Basin finance journal.

DiPasquale, D. & Wheaton, W. C. 1992. The markets for real estate assets and space: a conceptual framework. Journal of the American Real Estate & Urban Economics Association.

Diyan, M. Silva, B. N. & Han, K. 2020. A Multi-Objective Approach for Optimal Energy Management in Smart Home Using the Reinforcement Learning. Sensors.

Dufitinema, J. 2021. Stochastic volatility forecasting of the Finnish housing market. Applied economics, Vol. 53, p.98-114

Ehrig, T & Schmidt, J. 2021. Making biased but better predictions: The tradeoffs strategists face when they learn and use heuristics. Strategic Organization, 19(2), 263-284. https://doi.org/10.1177/1476127019869646

Einiö, M. Kaustia, M. & Puttonen, V. 2008. Price setting and the reluctance to realize losses in apartment markets. Journal of Economic Psychology.

Fan, G-Z. Ong, S. E. & Koh, H. C. 2006. Determinants of House Price: A Decision Tree Approach. Urban Studies.

Frost, J. n.d. Mean Squared Error (MSE). Statistics By Jim. Retrieved on 4.2.2023 from https://statisticsbyjim.com/regression/mean-squared-error-mse/

Fuerst, F. & Warren-Myers, G. 2018. Does voluntary disclosure create a green lemon problem? Energy-efficiency ratings and house prices. Energy Economics.

Gaetano, L. 2015. Real Estate Macroeconomics and the Four-Quadrant Model: DiPasquale-Wheaton-Colwell Meet Mortensen-Pissarides. Journal of Real Estate Practice and Education. Vol. 18, Iss. 1.

Garreta, R. & Moncecchi, G. 2013. Learning scikit-learn: Machine Learning in Python. Packt Publishing.

Genesove, D. & Mayer, C. 2001. Loss Aversion and Seller Behavior: Evidence from the Housing Market. The Quarterly journal of economics.

Grigsby, W. G. 1963. Housing Markets and Public Policy. University of Pennsylvania Press.

Grinblatt, M. & Keloharju, M. 2001. What Makes Investors Trade? Wiley.

Hagerlund, T. 2022. Kaupunkien ja kuntien lukumäärät ja väestötiedot. Kuntaliitto. Retrieved from https://www.kuntaliitto.fi/kuntaliitto/tietotuotteet-ja-palvelut/kaupunkien-ja-kuntien-lukumaarat-ja-vaestotiedot.

Hannonen, M. 2014. Urban Housing Policy Considerations: Perspectives from the Finnish Housing Market. Journal of heterodox economics.

Hastie, T.J. & Tibshirani, R. J. 2017. Generalized additive models. Routledge 2017.

Hawkins, D. M. 1980. Identification of Outliers. Springer Dordrecht.

He, X. Zhao, K. Chu, X. 2021. AutoML: A survey of the state-of-the-art. Elsevier. https://doi.org/10.1016/j.knosys.2020.106622

Henriksson, E. & Werlinder, K. 2021. Housing Price Prediction Countrywide Data – A comparison of XGBoost and Random Forest regressor models. Degree Project in Technology. KTH Royal Institute of Technology.

Hong, J. Choi, H. & Kim, W. 2020. A House Price Valuation Based on The Random Forest Approach: The Mass Appraisal of Resiential Property in South Korea. International Journal of Strategic Property Management.

Huang, S-C & Le, T-H. 2021. Principles and Labs for Deep Learning. Elsevier.

Huomo, M. Kannisto, O. 2022. Muutot kerrostaloista pientaloihin yleistyneet koronavuosina. Statistics Finland.

IBM. 2023. Adjusted R squared. Retrieved on 4.2.2023 from https://www.ibm.com/docs/fi/cognos-analytics/11.1.0?topic=terms-adjusted-r-squared.

Iwai, K. & Hamagami, T. 2022. A New XGBoost Inference with Boundary Conditions in Real Estate Price Prediction. IEEJ transactions on electrical and electronic engineering.

Jadhav, A. Pramod, D. & Ramanathan, K. 2019. Comparison of Performance of Data Imputation Methods for Numeric Dataset. Applied Artificial Intelligence.

Jaiswal, K. B. Patil, H. 2020. The Study Using Ensemble Learning for Recommending Better Future Investments. International Journal of Advanced Research in Computer Science.

James, G. Witten, D. Hastie, T. & Tibshirani, R. 2017. An Introduction to Statistical Learning with Applications in R. Springer.

Jha, S. B. Babiceanu, R. F. Pandey, V. Jha, R. K. 2020. Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study. Cornell University.

Jiang, Y. & Qiu, L. 2022. Empiricial study on the influencing factors of housing price --- Based on cross-section data of 31 provinces and cities in China. Procedia Computer Science.

Jim, C.Y. & Chen, W. Y. 2006. Impacts of urban environmental elements on residential housing prices in Guangzhou (China). Landscape and urban planning.

Kannisto, O. Korhonen, M. Rämö, A & Vuorio, E. 2020. Yli puolet viime vuonna myydyistä yksiöistä meni sijoittajille. Statistics Finland. Retrieved 17.4.2022 from https://www.stat.fi/tietotrendit/artikkelit/2020/yli-puolet-viime-vuonna-myydyista-yksiosta-meni-sijoittajille/

Kalliola, J. Kapočiūtė-Dzikienė, J. Damaševičius, R. 2021. Neural network hyperparameter optimization for prediction of real estate prices in Helsinki. PeerJ Comput. Sci. 7: e444 DOI 10.7717/peerj-cs.444

Kaustia, M. 2001. Disposition Effect. Behavioral Finance – Investors, Corporations and Markets. Wiley.

Kayakuş, M. Terzioğlu, M. & Yetiz, F. 2022. Forecasting housing prices in Turkey by machine learning methods.

Kiel, K. A. & Zabel, J. E. 2004. Location, Location, Location: The 3L Approach to House Price Determination. Bureau of the Census.

Kim, D. H. & Irakoze, A. 2023. Identifying Market Segment for the Assessment of a Price Premium for Green Certified Housing: A Cluster Analysis Approach. Sustainability. https://doi.org/10.3390/su15010507

Koengkan, M. & Fuinhas, J. A. 2022. Heterogeneous Effect of ”Eco-Friendly” Dwellings on Transaction Prices in Real Estate Market in Portugal. Energies (Basel)

Korstanje, J. 2021. Advanced Forecasting with Python: With State-Of-the-Art-Models Including LSTMs, Facebook’s Prophet, and Amazon’s DeepAR. Berkeley, CA: Apress L. P. 2021.

Krizhevsky, A. Sutskever, I. & Hinton, GE. 2012. ImageNet classification with deep convolutional neural networks. In: Proceedings of the NIPS, Lake Tahoe, CA, USA1097–1105.

KTI Kiinteistötieto Oy. 2022. The Finnish Property Market. Retrieved from https://view.taiqa.com/kti/finnish-property-market-2022%20

Kumar, G. K. Rani, D. M. Koppula, N. & Ashraf, S. 2021. Prediction of House Price Using Machine Learning Algorithms. 5th International Conference on Trends in Electronics and Informatics.

Laaksonen, K. 2022. Machine Learning in House Price Prediction – Case Study of Hous Price Prediction in the Helsinki Metropolitan Area. Master’s thesis. Aalto University.

Leech, D. Campos, E. 2001. Is Comprehensive Education Really Free? A Case Study of The Effects of Secondary School Admissions Policies On House Prices In One Local Area. IDEAS Working Paper Series from RePEc.

Liu, G. 2022. Research on Prediction and Analysis of Real Estate Market Based on the Multiple Linear Regression Model. Scientific Programming.

Louati, A. Lahyani, R. Aldaej, A. Aldumaykhi, A. & Otai, S. 2022. Price forecasting for real estate using machine learning: A case study on Riyadh city. Concurrency and Computation.

Manasa, J. Gupta, R. & Narahari, N. S. 2020. Machine Learning based Predicting House Prices using Regression Techniques. Proceedins of the Second International Conference on Innovative Mechanisms for Industry Applications.

Martin, A. & Ryan-Collins, J. 2016. The Financialisation of UK Homes – The housing crisis, land and the banks. New Economics Foundation.

Matplotlib documentation. N.D. Retrieved on 19.2.2023 from https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.boxplot.html

Mora-Garcia, R-T. Cespedes-Lopez, M-F. & Perez-Sanchez, V. R. 2022. Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times. Land.

Murphy, K. 2012. Machine learning a probabilistic perspective. MIT Press.

Nedjanti, G. L. Schneider, T. Hall, M. G. & Cawley, N. 2017. Machine learning based compartment models with permeability for white matter microstructure imaging. NeuroImage.

Ntantamis, C. 2010. Detecting Housing Submarkets using Unsupervised Learning of Finite Mixture Models. EconPapers.

Official Statistics of Finland (OSF): Dwellings and housing conditions [e-publication]. ISSN=1798-6761. Overview 2020, 1. Dwelling stock 2020. Helsinki: Statistics Finland [referred: 17.4.2022].
Access method: http://www.stat.fi/til/asas/2020/01/asas_2020_01_2021-10-14_kat_001_en.html

Oikarinen, E. 2007. Studies on housing price dynamics. Doctoral Thesis. Turku School of Economics. Retrieved from: https://urn.fi/URN:ISBN:978-951-564-507-4.

Oikarinen, E. & Engblom, J. 2016. Differences in housing price dynamics across cities: A comparison of different panel model specifications. Urban studies.

Osborne, J. W. & Overbay, A. 2004. The power of outliers (and why researchers should ALWAYS check for them)," Practical Assessment, Research, and Evaluation: Vol. 9, Article 6. DOI: https://doi.org/10.7275/qf69-7k43

Osland, L & Pryce, G. 2012. Housing Prices and Multiple Employment Nodes: Is the Relationship Nonmonotonic? Taylor & Francis Online.
https://doi.org/10.1080/02673037.2012.728571

Pakarinen, S. 2018. Semiparametric Efficiency Analysis of Housing Markets. Doctoral Thesis. Aalto University.

Park, J. Y. Dougherty, T. Fritz, H. & Nagy, Z. 2019. LightLearn: An adaptive and occupant centered controller for lighting based on reinforcement learning. Building and environment.

Peng, Z. Huang, Q. & Han, Y. 2019. Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGBoost Algorithm.

Płoński. P. 2020. XGBoost Feature Importance Computed in 3 Ways with Python. Mljar. Retrevied on 2.4.2023 from https://mljar.com/blog/feature-importance-xgboost/.

Rogel-Salazar, J. 2018. Data Science and Analytics with Python. Chapman and Hall.

Ryan-Collins, J. Lloyd, T. Macfarlane, L. & Muellbauer, J. 2017. Rethinking the Economics of Land and Housing. Bloomsbury Academic & Professional.

Saastamoinen, U. Vikström, S. Helminen, V. Lyytimäki, J. Nurmio, K. Nyberg, E. & Rantala, S. 2022. The limits of spatial data? Sense-making within the development and different uses of Finnish urban-rural classification. Land use policy, Vol. 120.

Sawant, R. Jangid, Y. Tiwari, T. Jain, S. & Gupta, A. 2018. Comprehensive Analysis of Housing Price Prediction in Pune Using Multi-Featured Random Forest Approach. 2018

Fourth International Conference on Computing Communication Control and Automation.

Singh, K. & Upadhyaya, S. 2012. Outlier Detection: Applications and Techniques. IJCSI International Journal of Computer Science Issues.

Stein, J. C. 1995. Prices and Trading Volume in the Housing Market: A Model with Down-Payment Effects. The Quarterly journal of economics.

Subasi, A. 2020. Practical Machine Learning for Data Analysis Using Python. Elsevier.

Suomen virallinen tilasto (SVT): Osakeasuntojen hinnat [verkkojulkaisu]. ISSN=2323-878X. Lokakuu 2021. Helsinki: Tilastokeskus [viitattu: 9.4.2022]. Saantitapa: http://www.stat.fi/til/ashi/2021/10/ashi_2021_10_2021-11-30_tie_001_fi.html

Statistics Finland. N.D. Paavo postal code area statistics. Retrieved from https://www.stat.fi/tup/paavo/index_en.html.

The Housing Finance and Development Centre of Finland (ARA). N.D. Asuntojen.hintatiedot.fi.

Truong, Q. Nquyen, M. Dang, H. & Mei, B. 2019. Housing Price Prediction via Improved Machine Learning Techniques. Elsevier.

Wang, Y. Li, Y. Sony, Y. & Rong, X. 2020. The influence of the activation function in a convlotuion neural network model of facial expression recognition. Applied Sciences. 10(5):1897 DOI 10.3390/app10051897

Wei-Meng, L. 2019. Python machine learning. John Wiley & Sons. https://ebookcentral-proquest-com.ezproxy.vasa.abo.fi/lib/abo-ebooks/reader.action?docID=5747364

Whieldon, L. & Ashqar, H. I. 2022. Predicting residential property value: a comparison of multiple regression techniques. SN Business & Economics.

Wu, Y. Wei, Y. D. & Li, H. 2019. Analyzing Spatial Heterogeneity of Housing Prices Using Large Datasets. Applied Spatial Analysis and Policy.

Xu, L. & Li, Z. 2021. A New Appraisal Model of Second-Hand Housing Prices in China’s First-Tier Cities Based on Machine Learning Algorithms. Computational economics.

Yang, J. 2020. Academic dissertation: Outlier Detection Techniques. University of Eastern Finland.

Zhang, Q. 2021. Housing Price Prediction Based on Multiple Linear Regression. Scientific programming.

Zhang, Y. Huang, J. Zhang, J. Liu, S. & Shorman, S. 2022. Analysis and prediction of second-hand house price based on random forest. Applied Mathematics and Nonlinear Sciences, vol 7, no.1, 2022, pp.27-42.

Zhang, Z. 2021. Decision Trees for Objective House Price Prediction. 3rd International Conference on Machine Learning, Big Data and Business Intelligence.