University of Tartu

Faculty of Social science

Narva College

Information systems development

Kristina Kudryavtseva

ESTONIAN REAL ESTATE PRICES PREDICTION USING MACHINE LEARNING

Diploma thesis

Supervisor: assistant Andre Sääsk

Narva 2023

Olen koostanud töö iseseisvalt. Kõik töö koostamisel kasutatud teiste autorite tööd, põhimõttelised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on viidatud.

.....

Töö autori allkiri ja kuupäev

Non-exclusive license to reproduce thesis and make thesis public

I, Kristina Kudryavtseva (date of birth: 27.09.1998),

1. herewith grant the University of Tartu a free permit (non-exclusive license) to:

1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

ESTONIAN REAL ESTATE PRICES PREDICTION USING MACHINE LEARNING,

supervised by Andre Sääsk,

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive license does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Narva, 14.05.1998

ABSTRACT

The goal of this study is to use a variety of machine learning approaches to build an accurate and dependable model for predicting real estate prices. The study analyzes and identifies the primary elements that influence the values of properties in various regions using data from a variety of sources, including real estate listing websites and government databases. In order to enhance the performance of the models, the study also takes into account feature engineering approaches such, feature scaling and dimensionality reduction. The study's findings show that the Gradient boosting regression works better than other models, with an coefficient of determination 0.89, in predicting real estate prices. The study also reveals how crucial feature choice, data quality, and model tweaking are for enhancing model performance.

TABLE OF CONTENTS

LIST OF TERMS AND ABBREVIATIONS

ML – Machine Learning

GDP – Gross domestic product

ARIMA – Autoregressive integrated moving average model. It is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends. [1]

UI – User Interface

[1] https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp

1. INTRODUCTION

At the moment, there are many different studies and scientific papers in the field of real estate valuation and price forecasting around the world. For a number of parties, including buyers, sellers, real estate brokers, and investors, an accurate projection of real estate prices is crucial. Real estate agents can utilize price prediction models to advise clients on property pricing strategies, and accurate price prediction can assist buyers, investors, and investors in making informed decisions about buying or selling properties. Most of the studies are based on a standard approach of predicting real estate prices. They are based on expert knowledge and subjective evaluations of numerous aspects of a property's size, location, number of living rooms, the year of construction, and so on. Such data is great for predicting the price of a particular apartment, with certain parameters and location, but Estonian open data warehouse does not have enough data on completed transactions. So, in this thesis the author uses data's power to find trends and connections between different economic indicators and real estate prices. I want to investigate the possibility of predicting the real estate prices in Estonia using GDP, population, etc.

1.1. PROLBEM OVERVIEW

The real estate prices in Estonia are skyrocketing. It is difficult to decide whether now is a suitable time to invest in new housing or not. The problem that I want to reveal in my work is the unpredictability of the real estate market and check whether real estate prices can be predicted using data analysis and machine learning or not. Accurate property assessment can assist buyers and sellers in making accurate choices and prevent them from overpaying or underpricing a property. Also, investors can identify potential market downturns and adjust their investment strategies accordingly by predicting future property prices.

1.2. THESIS GOAL

The aim of this study is to collect and analyze various economic indicators (available on Estonian local websites) and to use a variety of machine learning approaches to build an accurate and dependable model for predicting real estate prices. The study analyzes and identifies the primary elements that influence the values of properties in various regions using data from a variety of sources, including real estate listing websites and government databases. In order to anticipate real estate prices, the study investigates various regression models, including linear regression, decision

tree regression, random forest regression, gradient boosting, CatBoost regression and support vector machines (SVMs). In order to enhance the performance of the models, the study also takes into account feature engineering approaches such one-hot encoding, feature scaling, and dimensionality reduction and for forecasting financial variables is used ARIMA models. Deploy the database where the data will be stored and converted into a flat table for further work with algorithms. And as a last step is creating a user-friendly web page with data analyzes and form with request where user can see the result by their parameters.

1.3. THESIS TASKS

To achieve this goal, the following steps must be performed:

- The study of the theoretical Data Science foundations and ready-made research on the topic of real estate price prediction.
- Research and choose the right tools for practical part.
- Data collecting, cleaning, and exploring.
- Database deploying.
- Data modeling:
    - Select modeling techniques depending on the dataset and expected results.
    - Generate test design.
    - Build and assess models.
- Results evaluation and review of processes.
- Deploy the project.

1.4. OUTLINE

The first chapter is an introduction to thesis. In this chapter, the author describes the problem and goal of this study and tasks to achieve this aim.

The second chapter is an overview of the theoretical background of machine learning models and data processing methodologies which was used in this thesis.

The third chapter includes information about used technologies and tools.

In the fourth chapter of thesis, the author describes the practical part of work:

- Data searching and collecting.
- Develop a database for data storing.
- Writing scripts for data processing automatization.
- Data analysis and preparation of data set for machine learning models
- ML models building, testing and comparison.
- Develop a simple application with UI for working with model.

In the fifth chapter, the author presents results of different Machine Learning models and summarizes quality of data and models.

The sixth chapter is an overview of the future development possibilities of the data features and ML models and data collecting and list of functionalities that can be added into the web application in future.

2. THEORETICAL BACKGROUND

2.1. REAL ESTATE ESTIMATION

According to recent data, the real estate market in Estonia has been experiencing steady growth in recent years, with prices rising across the country. However, there are certain areas that have seen higher levels of growth than others. For example, the capital city of Tallinn and its surrounding areas have seen the most significant increase in property prices, while other regions have experienced more moderate growth.

Experts predict that the Estonian real estate market will continue to grow over the next few years, but at a slower pace than in previous years. This is due in part to government policies aimed at preventing an overheated market and ensuring sustainable growth.

Property prices in Estonia very depending on the location and type of property. Generally, prices in Tallinn and its surrounding areas are higher than in other parts of the country. Like an example, according to data from the Estonian Land Board, the average price of a square meter of residential real estate in Tallinn was €2,312 in Q4 2021. It grew up to €2,965 in February 2023, up 9 percent on year. The average price per square meter came to €2,458 in Tartu for an annual growth of 22 percent. This was €2,345 and roughly 20 percent for Pärnu.

Rental market: The rental market in Estonia has also been growing in recent years, with high demand for rental properties in Tallinn and other urban areas. According to data from the Estonian Statistical Office, the average monthly rent for a one-room apartment in Tallinn was €492 in Q4 2021 and grew to €690 for a furnished studio in an expensive area in 2023.

2.2. PREDICTIVE MODELS

Predictive modeling is a tool which is used in predictive analytics. It refers to the process of using mathematical statistics and computational methods to develop predictive models that use historical data and already known results for training and predicting results for new inputs. Modeling provides

results in the form of predictions representing the probability of a target variable based on the assumed significance of a set of input variables.

We can divide predictive models into 2 types: supervised and unsupervised learning. Supervised learning models have a specified output which can predict class membership (classification models) or number (regression models). Usually, the classification model results have the binary form 1 or 0 (true or false). Unsupervised learning models do not have any target variable and are often used during exploratory data analysis to identify patterns or natural groupings in the data.

Based on this description, it can be understood that for the purposes of this thesis, supervised learning is better suited, more precisely, regression models.

Predictive algorithms used in this thesis:

- Linear regression
- Decision tree regression
- Random forest regression
- Gradient boosting
- CatBoost regression
- Support Vector Machines

Also, for forecasting is used time series model called ARIMA (based on ARMA).

2.2.1. Linear regression

Linear regression is very popular methods in statistics. It supposes that the input features and the target variable (i.e., the price) have a linear relationship. In order to create predictions based on fresh inputs, the model calculates the parameters of a linear equation that best fits the data. Regression analysis is a reliable method of identifying which variables have an impact on a topic of interest. The process of performing a regression allows you to confidently determine which factors matter most, which factors can be ignored, and how these factors influence each other.[2]

[2] https://www.alchemer.com/resources/blog/regression-analysis/

Simple linear regression:

$y = B_0 + B_1 * x$

When training the model, we are given:

- x – input training data
- y – labels to data

During model training, the model obtains the best regression line by finding the best values for B_1 (coefficient of x) and B_0 (intercept). Once we find the best θ1 and θ2 values, we get the best fit line. Once we find the best values for a and b, we can finally use our model to make predictions.

2.2.2. Decision tree

Decision trees are a widely used machine learning algorithm that can be used for both regression and classification tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.[3] In this study is used CART (Classification and Regression Tree). In case of regression tree, the value obtained by terminal nodes in the training data is the mean response of observation falling in that region. Thus, if an unseen data observation falls in that region, we'll make its prediction with mean value.[4]

The main stages of construction:

- Selection of the attribute by which splitting will be performed in this node (split attribute).
- Selection of the learning stop criterion.
- Selection of the method of cutting off branches (simplification).
- Estimation of the accuracy of the constructed tree.

Useful parameters for tree modeling:

[3] https://scikit-learn.org/stable/modules/tree.html
[4] https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/

Information gain measures the reduction of uncertainty given some feature and it is also a deciding factor for which attribute should be selected as a decision node or root node.[5]

Classification criteria:

- **Gini**
- Log loss or **entropy**

Regression criteria:

- Mean Squared Error
- Half Poisson deviance

2.2.3. Random forest regression

The Random Forest algorithm combines ensemble learning techniques with the decision tree framework to generate numerous randomly selected decision trees from the input. The program then averages the results to provide a new result that frequently produces accurate predictions and classifications. A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.[6] The ability of the Random Forest Algorithm to handle data sets with both continuous variables, as in the case of regression, and categorical variables, as in the case of classification, is one of its most significant features.

2.2.4. Gradient boosting

Gradient boosting is a method standing out for its prediction speed and accuracy, particularly with large and complex datasets. The primary concept underlying this algorithm is to build models in succession while attempting to minimize the flaws of the prior model. By developing a new model on the errors or residuals of the prior one, this is accomplished.

[5] https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/#What_is_a_Decision_Tree?
[6] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html#:~:text=A%20random%20forest%20regressor.,accuracy%20and%20control%20over%2Dfitting.

The objective in boosting is to minimize loss function by adding weak learners using gradient descent. Since it is based on a loss function, we will have various loss functions for regression problems, such as Mean Squared Error (MSE), and for classification problems, such as log-likelihood.

2.2.5. Neural Network

As mentioned above, a neural network is used for predictions. Neural network is a subset of machine learning which are roughly modeled the human brain, which are designed to recognize patterns. The patterns they recognize are numeric contained in vectors into which time series data must be translated. If neural network has enough right data, deep learning is able to establish correlations between present events and future events. Deep learning doesn't necessarily care about time, or the fact that something hasn't happened yet. Given a time series, deep learning may read a string of number and predict the number most likely to occur next.[7]

NN components:

- Input layer
- Processing layers
- Output layer

The inputs may be weighted based on various criteria. Within the processing layer, which is hidden from view, there are nodes and connections between these nodes, meant to be analogous to the neurons and synapses in an animal brain.[8]

2.2.6. Support Vector Machines

Support Vector Machines was made for classification, but it is functions was extended to the task of regression and forecast as well. SVR (Support Vector Regression) is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same

[7] https://wiki.pathmind.com/neural-network
[8] https://www.investopedia.com/terms/n/neuralnetwork.asp

principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.[9]

The SVR strives to fit the best line within a threshold value, as opposed to other Regression models that aim to reduce the error between the real and projected value. The gap between the hyperplane and boundary line is the threshold value. Since SVR's fit time complexity is more than quadratic with the number of samples, it is challenging to scale datasets with more than a few tens of thousands of samples.

2.2.7. CatBoost regression

A machine learning algorithm called CatBoost regression was created specifically for regression tasks. It is recognized for its ability to handle categorical data without the need for intensive preprocessing and is an extension of gradient boosting techniques. To produce precise predictions on numerical target variables, CatBoost combines gradient boosting and decision trees. In order to enhance performance, it also uses cutting-edge strategies including ordered boosting and feature combinations.

CatBoost offers the possibility to extract Variable Importance Plots. Hence, a Variable Importance Plot could reveal underlying data structures that might not be visible to the human eye.[10]

2.2.8. Time Series Analysis and Forecasting

A time series is simply a series of data points ordered in time. In a time series, time is often the independent variable and the goal is usually to make a forecast for the future.[11] When working with time series data, there are several aspects that need to be taken into account:

- Autocorrelation – is the similarity between observations as a function of the time lag between them.

[9] https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0#:~:text=Support%20Vector%20Regression%20is%20a,the%20maximum%20number%20of%20points.

[10] https://towardsdatascience.com/catboost-regression-in-6-minutes-3487f3e5b329

[11] https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775

- Seasonality - data change that occurs at regular intervals.
- Stationarity is indicated by the invariance of the mean and variance of the data.

In this thesis we will deeply study ARIMA (Autoregressive integrated moving average) model. ARIMA is a forecasting algorithm based on the idea that the information in the past values can be used to predict the future values without any additional values.

The first part of building an ARIMA model is to make the time series stationary, because AR in ARIMA means a linear regression model that uses its own lags as predictors.[12] This is a regression where Y_t depends only on its own lags.

Y_t function of the lags:

$$Y_t = a + B_1Y_{1-1} + B_2Y_{t-2} + .. + B_pY_{t-p} + e_1$$

Where,

- Y_{t-1} is the lag_1 of the series.
- B_1 is the coefficient of lag_1 (estimated by the model)
- a is the intercept term (estimated by the model)

And the second model in ARIMA is Moving Average model, where Y_t depends only on the lagged forecasts errors.

$$Y_t = a + e_t + \varphi_1 e_{t-1} + \varphi_2 e_{t-2} + .. + \varphi_q e_{t-q}$$

Where the e are the errors of the AR models of the respective lags.
ARIMA is a model where the time series has been converted to stationary and you combine the AR and MA terms.

[12] https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

3. USED TECHNOLOGIES AND TOOLS

3.1. Data preprocessing methods

Data preprocessing is a procedure where we clean and solve most of the issues in the data.

Data preprocessing methods:

- **Collecting**: data collecting from various sources, such as databases, web resources, files, etc. Ensure that you have sufficient data that is relevant to your problem statement.
    - Since the data is collected from various sources with different conditions, first of all, data is needed to be clean and merged into one flat table. The main part of manipulations with data were performed in the MS Excel environment.
- **Exploration**: for a better understanding of the data, we can perform the data exploration, such as identifying missing values, outliers, and anomalies. This can help to decide how to handle these problems.
    - Anaconda navigator with Jupyter Notebook is used for detailed analysis and study of data reduced to a general form. This is a classic environment for working with data using Python language and powerful libraries such as NumPy, pandas, seaborn, matplotlib, pandas-profiling (reporting function).
- **Cleaning**: is used to handle missing values, outliers, and anomalies. There are lots of different methods to process missing values, such as filling them with the mean or median value or removing the rows with missing values altogether. Analogously, outliers can be removed or replaced with a more appropriate value.
- **Transformation**: data transformation is used to make it more suitable for Machine Learning algorithms. This may include scaling, normalization, or encoding categorical variables.
- **Splitting**: splitting the data into training and testing sets.
- **Feature Selection**: selection of the most relevant features for your problem statement. Feature selection can help to reduce the complexity of the model and improve its accuracy.
- **Model Development**: Developing the Machine Learning model by using the cleaned and transformed information and data. This could involve selecting the appropriate algorithm,

training the model, and optimizing its performance. For data modeling, training, testing and evaluation is used library sklearn.

3.2. Data scraping

The process of automatically obtaining data from webpages is called web scraping. It entails using software tools to extract data from websites, including text, photographs, and videos, and then storing it in a structured format for later processing or analysis.

Scraping will be used to collect data about real estate offers at the current moment. Since the site we have chosen has a dynamic pages, a script written in python and selenium libraries (including web driver) will be used to collect such kind of data as price, county, number of rooms, area, etc.

3.3. Model deploying

Deployed locally on personal PC.

3.4. Simple web page

Flask is selected as the main back-end framework for REST API, which is written in Python. It has multiple modules that make it easier to write web applications without thinking about thread management, protocol management, etc.

Bootstrap will be used for designing simple front-end for our web page.

4. PRACTICAL PART

This chapter will describe the practical part of the work.

4.1. DATA PREPARATION

The main part of creating and using any kind of Machine Learning instruments is cleaning and preparing data. Also, those are more critical steps in any machine learning process. This chapter will describe the practical part of the work.

4.1.1 Data collecting

We will use three main local resources with data:

1. Estonian Land Board, transactions database [13]
2. Statistical database [14]
3. Kinnisvara24.ee

From Estonian Land Board, transactions database was taken data with parameters:

- Type of publication: Price statistics
- Report: Transactions with residential apartments
- County: All
- Several time periods:
    - select years: 2005-2023
    - select quarter: all

Reports used from statistical database:

- RAA0012: GROSS DOMESTIC PRODUCT AND GROSS NATIONAL INCOME (ESA 2010) (QUARTERS)

[13] https://www.maaamet.ee/kinnisvara/htraru/
[14] https://andmed.stat.ee/et/stat

- RV032: BIRTHS, DEATHS AND NATURAL INCREASE BY SEX AND COUNTY
- RAA0014: DISPOSABLE INCOME, SAVING AND NET LENDING/BORROWING AT CURRENT PRICES (ESA 2010) by Year, Quarter and Indicator
- RAA0050: GROSS DOMESTIC PRODUCT BY COUNTY (ESA 2010)

All reports were downloaded manually from sites in *.xlsx* format. All data was formatted and brought to a single view. Since there is no report in the public domain where is shown GDP by counties and quarters, 2 reports were used. Based on report RAA0050, a proportion was calculated to split the quarterly results by county. On Picture 1 is shown part of proportion calculation.

Sum of GDP at current prices, million euros	Column Labels		
Row Labels	2021	2020	2019
Harju county	63.11%	62.73%	61.86%
Hiiu county	0.42%	0.42%	0.41%
Ida-Viru county	5.83%	5.84%	6.55%
Järva county	1.57%	1.60%	1.69%
Jõgeva county	1.24%	1.28%	1.23%
Lääne county	0.92%	0.95%	0.95%
Lääne-Viru county	2.82%	2.86%	3.07%
Pärnu county	4.11%	4.14%	4.31%
Põlva county	0.88%	0.90%	0.93%
Rapla county	1.45%	1.47%	1.45%
Saare county	1.47%	1.49%	1.49%
Tartu county	11.22%	11.28%	10.99%
Valga county	1.17%	1.20%	1.15%
Viljandi county	2.33%	2.36%	2.43%
Võru county	1.47%	1.48%	1.49%
Grand Total	**100.00%**	**100.00%**	**100.00%**

Picture 1. Proportion calculation in excel

From *Kinnisvara24.ee* was scraped data about currently available offers. Script is written in Python language with the use of Selenium libraries including web driver. First of all, script collects all available data in page one by one and save link to each sale announcement. After all pages is checked, script start to analyze every announcement page separately for handling and collecting details about apartment (price, city, area, floor, number of rooms, building year, building type,

condition, etc). When driver finish work it save all data in the *.csv* file with certain name, for example '*Scrapping_results mm-yyyy.cvs'*.

	Name	County	Price	Floor	Area	Rooms	Build_year	Build_type	E_class	Condition
0	Koskla tn 20, Lilleküla, Kristiine	Harju maakond	209460.0	3.0	66.3	2.0	1960.0	Kivimaja	D	Renoveeritud
1	A. Weizenbergi tn 12a, Kadriorg, Kesklinn	Harju maakond	400000.0	2.0	89.1	4.0	1912.0	Puumaja	E	NaN
2	Liivalaia tn 21, Tatari, Kesklinn	Harju maakond	240000.0	4.0	58.0	2.0	2005.0	Paneelmaja	E	NaN
3	Narva mnt 40, Raua, Kesklinn	Harju maakond	225000.0	3.0	44.7	2.0	2018.0	Kivimaja	B	NaN
4	Roosikrantsi tn 15, Kesklinn	Harju maakond	260000.0	5.0	114.5	3.0	1912.0	Kivimaja	NaN	NaN
5	Kivila tn 34, Mustakivi, Lasnamäe	Harju maakond	139000.0	9.0	60.5	3.0	1980.0	NaN	D	Renoveeritud
6	Pärnu mnt 127, Kitseküla, Kesklinn	Harju maakond	133900.0	2.0	36.6	2.0	NaN	Kivimaja	F	Heas korras
7	Sõpruse pst 228, Mustamäe, Mustamäe	Harju maakond	145000.0	5.0	60.3	4.0	1965.0	Paneelmaja	E	NaN
8	Madara tn 1-1, Lilleküla, Kristiine	Harju maakond	274000.0	3.0	62.7	3.0	2022.0	Kivimaja	B	Uus
9	Vasara tn 15-12	Harju maakond	132000.0	1.0	39.2	2.0	2022.0	Kivimaja	A	Uus

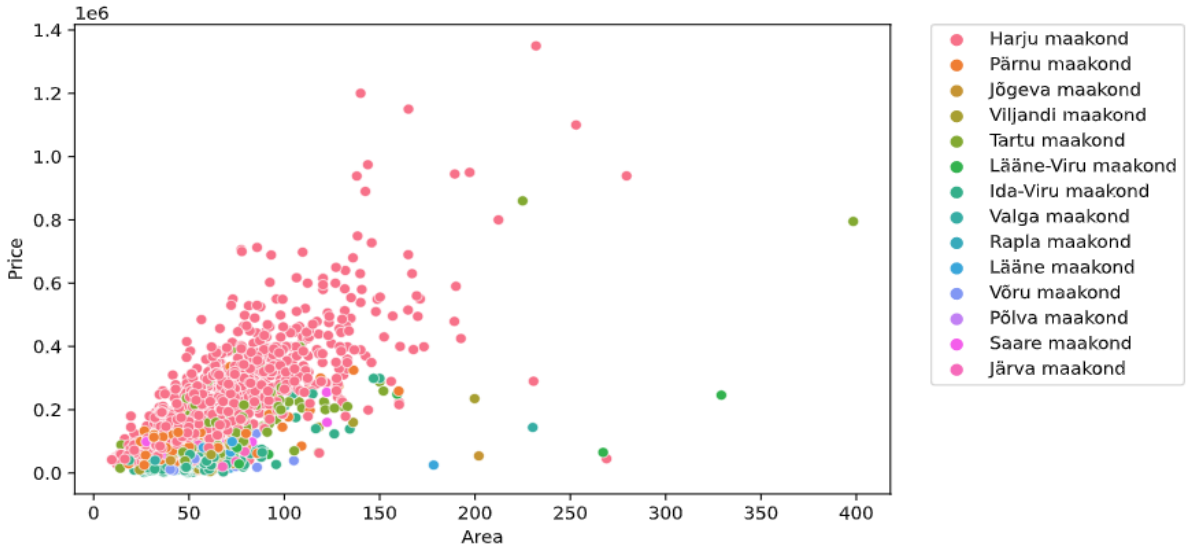
Picture 2. Example of Data collecting results

4.1.2 Data exploration and analysis

Exploring it is the next step. In order to obtain insights and spot trends, exploratory data analysis entails visualizing and summarizing the data. This stage helps in gaining an understanding of the data structure and content and could provide useful details regarding the connections between the various elements of the data.

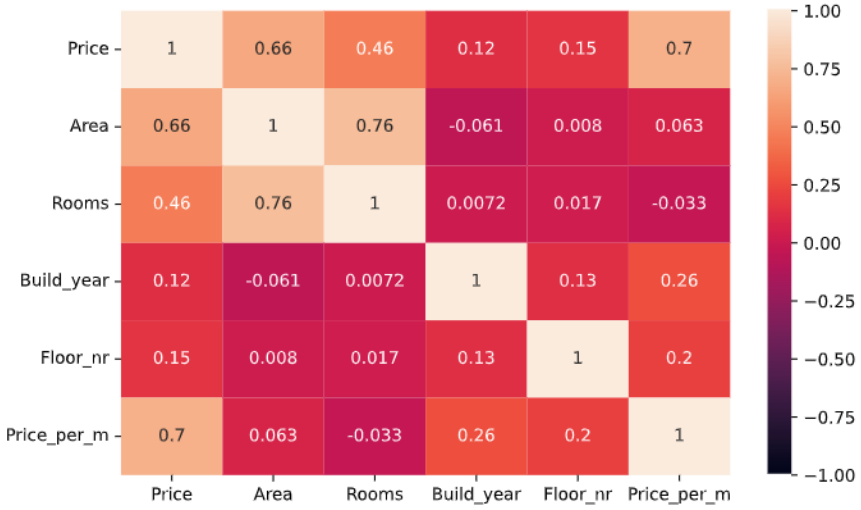
As a results of data exploration we can say what the most of the sale announcement is placed in Harju county. It is shown on picture below. Also, in this picture you can see the presence of exceptions, single ads that stand out from the overall picture. To improve the model, these data should be excluded from the sample.

In particular, we may see that the number of ads in certain regions is insignificant, which will make it more difficult to predict a more accurate price. Therefore, to create a model, 4 cities were selected (Tallinn, Tartu, Pärnu, Narva) from the most popular counties (Harju county, Tartu county, Pärnu county and Ida-Virumaa county). You can find a detailed analysis in the attached "*Primary data analysis.ipynb*" file, which is made on the example of data collected in the second quarter of 2022.



Picture 3. Scatter plot of collected data by county.

From picture 4 we can see the correlation of various parameters from each other. The highest dependencies explored between the number of rooms and area. Otherwise between price and area.



Picture 4. Correlation of variables.

4.1.3 Data cleaning

After collecting the data, it is important to clean it to get rid of any discrepancies, errors, or missing values. This stage includes locating and managing any faults or irregularities in the data as well as making sure that the data is accurate, consistent and reliable.

Data cleaning includes:

- Identify missing values and fill in the blanks if it is possible or remove data with nulls.
- Handle outliers. Data points identified as outliers differ significantly from other data points in the dataset. For example, in this dataset it can be offers with extremely high prices. As you can see individual cases with prices higher than 1 million EUR in picture number 2.
- Removing duplicates. Duplicates can be identified by rows checking that have identical values for all columns in the dataset. It could be removed by keeping only one copy of each row or by merging the duplicate rows.
- Normalize data. The process of normalizing data is adjusting it to have a constant scale or range. Some machine learning algorithms that are sensitive to the scale of the input data may find this to be essential. Standardization, min-max scaling, and logarithmic scaling are methods for normalization.
- Encode categorical variables. Categorical variables, like gender or color, are variables that can only have a specific set of values. It is necessary to encode these variables so that the algorithm used for machine learning can understand them. One-hot encoding, label encoding, and binary encoding are all methods for encoding categorical information.

4.1.4 Data Transformation

A key phase in preparing data for machine learning is data transformation. The procedure entails formatting the input data in a way that machine learning algorithms are able to use it. In picture 5 is shown how population and GDP was transformed with *MinMaxScaler* function from *sklearn.preprocessing* library. It was done to reduce the error of the model.

	Price	Area	Rooms	new	old	Narva	Pärnu	Tallinn	Tartu	Index	Population	GDP
5722	443758	102.7	4	1	0	0	0	1	0	5.48302	1.000000	1.000000
11824	76101	36.4	2	0	1	0	0	0	1	5.48302	0.137887	0.173002
1656	316221	74.3	3	1	0	0	0	1	0	5.54779	0.992679	0.885695
4157	84433	18.8	1	0	1	0	0	1	0	5.57640	0.992679	0.958953
5294	142098	43.5	2	0	1	0	0	1	0	6.06873	0.992679	0.966277

Picture 5. Final dataset

4.1.5 Time series data forecasting with ARIMA model

Details of forecasting is shown in attached files:

- Forecast for GDP.ipynb
- Forecast for price index.ipynb
- Forecast for population.ipynb

Below you can see examples of results of forecasting.

4.1.5.1 Forecast for GDP

	Quarter	GDP	Harju	Ida	Tartu	Pärnu
495	2025-07-01	442.138337	0	0	0	1
496	2025-10-01	443.349701	0	0	0	1
497	2026-01-01	455.313737	0	0	0	1
498	2026-04-01	443.750289	0	0	0	1
499	2026-07-01	464.838679	0	0	0	1

Picture 6. Forecast for GDP.

4.1.5.2 Forecast for price index

Forecast based on quarterly report from eStat open warehouse.

	Quarter	Index
94	2027-01-01	5.964660
95	2027-04-01	6.096105
96	2027-07-01	6.349820
97	2027-10-01	6.596072
98	2028-01-01	6.763094

Picture 7. Forecast for price index.

4.1.5.3 Forecast for population

	Year	Population	Ida	Pärnu	Harju	Tartu
127	2028-01-01	165915.0	0	0	0	1
128	2029-01-01	167705.0	0	0	0	1
129	2030-01-01	169629.0	0	0	0	1
130	2031-01-01	171689.0	0	0	0	1
131	2032-01-01	173890.0	0	0	0	1

Picture 8. Population forecast example.

4.2 Data Modeling

For data modeling is used function from library sklear. Process and results is shown in attached file ‘ML models.ipynb’

Data split:

- Train_size – 70%
- Test size – 30%
- Random state – 100

4.2.1 Model estimation

- **Mean squared error (MSE):** The average of the squared discrepancies between the predicted and actual values is what this metric measures. Better performance is indicated by a lower MSE.
- **Coefficient of determination (R-squared):** This calculates the percentage of the dependent variable's variance that can be accounted for by the independent variables. When the R-squared value is 1, the model completely describes the data.
- **Mean absolute error (MAE):** This calculates the mean difference between the expected and actual values. A lower MAE denotes higher performance, similar to MSE.
- **Mean Absolute Percentage Error (MAPE) :** used to evaluate the accuracy of a regression model's predictions by measuring the average percentage difference between the predicted and actual values. A lower MAPE indicates better prediction accuracy

Main results of model training:

Model	**Model score (R-squared)**	**Mean Absolute Percentage Error (MAPE)**
Linear regression	0.7577	0.3127
Decision tree regression	0.8411	0.2077
Random forest regression	0.8598	0.2153
Gradient boosting	0.8881	0.2109
CatBoost regression	0.8328	0.2444
Support Vector Machines	-0.0267	1.1106

4.3 Model deploying

For deploying was selected Gradient boosting regression due to the best R-squared result. For final price prediction is used additional calculations which is done on back-end.

4.4 Final web app with front-end

Front end is presented as a simple one-pager with form where user can choose interested quarter, city, area, number of rooms and building state: new or old. When all data is putted into form user send request to the server where will be automatically added parameters based on quarter (price index, population and GDP) to the user input. And when data will be calculated with back-end, user will get approximate price, which is calculated according to the entered parameters.

5. FINAL RESULTS

The study's findings show that the Gradient boosting regression works better than other models, with an coefficient of determination (R-Squared) 0.89, in predicting real estate prices. The study also reveals how crucial feature choice, data quality, and model tweaking are for enhancing model performance.

6. FUTURE DEVELOPMENT

For further development, I want to go in more detail the possibilities of models settings and add deep learning (neutral networks) to the list of used models. Also, UI needs to be developed in future, when this model can be used to predictions in market. This UI can be shown like dashboard with data analysis, different forms and data inputs/outputs. And of course, I still feel data luck for accurate ML models work, so, it is necessary to continue data collecting, models monitoring and updating.

CONCLUSION

The aim of “ESTONIAN REAL ESTATE PRICES PREDICTION USING MACHINE LEARNING” study was to collect and analyze various economic indicators (available on Estonian local websites) and to use a variety of machine learning approaches to build an accurate and dependable model for predicting real estate prices. In this thesis the author uses data's power to find trends and connections between GDP, population and real estate prices. During work on this thesis was tried several models (Linear regression, Decision tree regression, Random forest regression, Gradient boosting, CatBoost regression, Support Vector Machines) and forecasting ARIMA model. The study's results show that the Gradient boosting regression works better than other models, with an coefficient of determination 0.89, in predicting real estate prices. The study also reveals how crucial feature choice, data quality, and model tweaking are for enhancing model performance. Summing up, I can say that a good result has been achieved in price prediction in separate cities in Estonia, but there are still a lot of details that can and should be improved.

KOKKUVÕTE

“EESTI KINNISVARAHINDADE ENNUSTAMINE MASINÕPE KASUTAMISEGA” uuringu töö eesmärk oli koguda ja analüüsida erinevaid majandusnäitajaid (saadaval Eesti kohalikel veebilehtedel) ning kasutada erinevaid masinõppe lähenemisviise, et ehitada üles täpne ja töökindel kinnisvarahindade prognoosimise mudel. Käesolevas lõputöös kasutab autor andmete jõudu, et leida trende ja seoseid SKT, rahvastiku ja kinnisvarahindade vahel. Töö käigus katsetati mitmeid mudeleid (Lineaarne regressioon, Otsustuspuu regressioon, Juhuslik metsa regressioon, Gradiendi võimendamine, CatBoost regressioon, Support Vector Machines) ja prognoosiv ARIMA mudel. Uuringu tulemused näitavad, et Gradient boosting regressioon toimib kinnisvarahindade ennustamisel paremini kui teised mudelid, mille determinatsioonikoefitsient on 0,89. Uuring paljastab ka, kui olulised on funktsioonide valik, andmete kvaliteet ja mudelite kohandamine mudeli jõudluse parandamiseks. Kokkuvõtvalt võin öelda, et Eesti eraldi linnades on hindade ennustamisel saavutatud hea tulemus, kuid detaile, mida saab ja tuleks parandada, on veel palju.

REFERENCES

1. Autoregressive Integrated Moving Average (ARIMA) Prediction Model. https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp (last viewed 14.05.2023)
2. What is Regression Analysis and Why Should I Use It? https://www.alchemer.com/resources/blog/regression-analysis/ (last viewed 14.05.2023)
3. Decision Trees. https://scikit-learn.org/stable/modules/tree.html (last viewed 14.05.2023)
4. Tree Based Algorithms: A Complete Tutorial from Scratch (in R & Python) https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/ (last viewed 14.05.2023)
5. Decision Tree Algorithm – A Complete Guide https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/#What_is_a_Decision_Tree? (last viewed 14.05.2023)
6. Random Forest Regressor documentation https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html#:~:text=A%20random%20forest%20regressor.,accuracy%20and%20control%20over%2Dfitting. (last viewed 14.05.2023)
7. A Beginner's Guide to Neural Networks and Deep Learning https://wiki.pathmind.com/neural-network (last viewed 14.05.2023)
8. What Is a Neural Network? https://www.investopedia.com/terms/n/neuralnetwork.asp (last viewed 14.05.2023)
9. Unlocking the True Power of Support Vector Regression https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0#:~:text=Support%20Vector%20Regression%20is%20a,the%20maximum%20number%20of%20points. (last viewed 14.05.2023)
10. CatBoost regression in 6 minutes https://towardsdatascience.com/catboost-regression-in-6-minutes-3487f3e5b329 (last viewed 14.05.2023)
11. The Complete Guide to Time Series Analysis and Forecasting https://towardsdatascience.com/the-complete-guide-to-time-series-analysis-and-forecasting-70d476bfe775 (last viewed 14.05.2023)
12. ARIMA Model – Complete Guide to Time Series Forecasting in Python https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/ (last viewed 14.05.2023)
13. Query of real property price statistics https://www.maaamet.ee/kinnisvara/htraru/FilterUI.aspx (last viewed 14.05.2023)
14. Estonian Statistical database https://andmed.stat.ee/et/stat (last viewed 14.05.2023)

APPENDICES

1. Appendix 1. Primary data analysis.ipynb
2. Appendix 2. Data final preparation.ipynb
3. Appendix 3. Forecast for GDP.ipynb
4. Appendix 4. Forecast for population.ipynb
5. Appendix 5. Forecast for price index.ipynb
6. Appendix 6. ML models.ipynb