Application of Machine Learning Methods for Forecasting Real Estate Value

Oleh Veres[1], Pavlo Ilchuk[1] and Olha Kots[1]

[1] *Lviv Polytechnic National University, Stepana Bandery str. 12, Lviv, 79013, Ukraine*

Abstract

The research is devoted to the study of models, methods and instruments of determining the value of real estate according to its physical parameters and location, as well as to the development of an algorithm for the application of machine learning methods in solving the task of estimating the value of real estate.

Correct assessment of real estate value plays a decisive role in the process of buying and selling. It is important to get an accurate real estate assessment in order to make informed decisions and avoid mistakes that can lead to financial losses. Valuation of real estate has historically been based primarily on manual data analysis and subjective evaluations, which often lead to errors and delays. The use of machine learning algorithms in solving this problem is more effective, as it has a number of advantages over the manual assessment method, namely: a high level of accuracy, exclusion of subjectivity and bias in assessments, time efficiency, cost reduction, use of geospatial data and justification of results.

The process of creating a machine learning model is conventionally divided into four stages, which include data collection, filtering, processing, addition, division into different samples and training the model based on this data. It was decided to use several regression algorithms at once to build machine learning models, in order to compare the results and choose the algorithm that is best suited to solving the task. By applying linear regression algorithms, decision trees, the nearest neighbor method, the support vector method, and random forest, the last one showed the best results. By manually selecting the hyperparameters for this algorithm, the mean absolute error of the predicted value was 8.49%, and the median was 1.9%.

For the implementation of the machine learning model and the development of an online system, a corresponding conceptual model was designed. Based on the tasks and functions that the program should perform, as well as the intended users and their capabilities, a diagram of use cases was built. This diagram serves to visualize the user and functional requirements of the system. To implement this system, it was decided to use a microservice architecture, which will ensure simplicity, flexibility and scalability.

Keywords

Analysis, Regression, Machine Learning, Forecasting the Real Estate Value, System Prototype

1. Introduction

There comes a moment in every person's life when the question of buying or selling real estate becomes relevant. Usually, such an event is extremely important, as it involves the involvement of significant funds and happens only a few times in a lifetime, so it requires a deliberate and reasoned approach. However, many people face limited access to objective information on real estate values and market trends.

Therefore, the presence of applications that will help facilitate, speed up, reduce the cost and rationalize this process is always relevant. On the other hand, the purchase of real estate can be considered as a type of investment, where the aforementioned speed, simplicity and rationality are also important. Real estate transactions involve the involvement of significant sums of money, so such decisions must be made on the basis of accurate information. Overvalued properties can sit on the market for a long time, while undervalued properties can result in significant losses for the seller. Therefore, it is extremely important to work with experienced and qualified specialists

COLINS-2024: 8th International Conference on Computational Linguistics and Intelligent Systems, April 12–13, 2024, Lviv, Ukraine

oleh.m.veres@lpnu.ua (O. Veres); pavlo.g.ilchuk@lpnu.ua (P. Ilchuk); olha.o.kots@lpnu.ua (O. Kots)

0000-0001-9149-4752 (O. Veres); 0000-0003-4636-2309 (P. Ilchuk); 0000-0001-7123-3635 (O. Kots)

who will be able to provide a reliable and impartial assessment of the property [1]. However, the cost of the services of such specialists can be significant. In addition, in the case of a purchase, the limitations of human capabilities should be taken into account, especially when it comes to a large amount of data. And for the buyer, it will mean lost opportunities. It should also be taken into account that the process of searching for real estate can take a considerable amount of time, and sometimes such a question can be urgent.

Today, there are various web platforms and services for real estate search, but they provide limited information, as they do not reflect its appraised value, and do not always include the correspondence between the price indicated by the seller and the real characteristics of the property. There are no adequate recommendation systems that would take into account and display comprehensive information for buyers and sellers.

Therefore, the development of an information system is relevant, as it will ensure the transparency of the real estate market, will make it possible to reduce costs for the services of realtors and real estate agents, and will enable users to more effectively make decisions related to buying or selling. *The goal of the research* – to develop a machine learning model that will provide users with up-to-date data on the value of real estate objects. This will allow sellers to determine the current value of the property depending on the parameters and location, in order to subsequently set the price at the time of sale. For buyers, this system will be relevant because it will make it possible to significantly speed up and simplify the process of searching for real estate by providing filtering, evaluation, recommendation, analysis and automatic notification tools.

2. Analysis of the current state and prospects in the field of research

2.1. An analytical review of the modern approach to real estate valuation using machine learning tools

The real estate assessment process plays a critical role in determining successful deals for both buyers and sellers. These assessments have historically relied primarily on manual data analysis and subjective assessments, which often lead to errors and delays.

However, the introduction of artificial intelligence marked the beginning of a period of change. This has impacted the way real estate is valued and opened the door to unprecedented levels of accuracy, efficiency and transparency. It is no longer necessary to rely entirely on the judgment of experts in the field, who often have to work with huge amounts of data and set complex criteria when determining real estate prices.

Artificial intelligence should be seen as a sophisticated researcher that examines a wide range of data sources, including real estate transaction history, tax assessments, similar properties, and regional market trends. Artificial intelligence adds a new dimension to the real estate valuation process by scrutinizing data and is able to provide information on fair market value, going beyond traditional methodologies in its analysis. The secret lies in the ability of artificial intelligence to detect subtle correlations and patterns in data, allowing for a more thorough and accurate assessment of property value.

Artificial intelligence algorithms act as virtual data miners, effectively sifting through data from a variety of sources. They work with a huge amount of data, while analyzing it much faster than any human can physically do, and turning it into useful information that will later serve as the basis for evaluations. Artificial intelligence provides a more comprehensive data-driven approach, ensuring that all relevant information is taken into account in the calculation process.

For artificial intelligence, a large amount of complex information is a playground, as this technology has the ability to recognize patterns, which brings to life hidden connections and trends that may be invisible to the human eye. Artificial intelligence enhances its valuation method by recognizing trends that can cross multiple dimensions, leading to a more thorough and detailed investigation of property value. This approach to pattern detection provides the evaluation process with a high level of accuracy and data insight that could hardly be achieved by conventional methods [2].

It should also be noted that the dynamics of the real estate market require the ability to adapt in real time, which artificial intelligence skillfully performs. Traditional valuation methods often rely on out-of-date data and are unable to account for market changes that occur rapidly. In contrast, artificial intelligence is able to smoothly incorporate changes in real time, which determines its absolute advantage.

As a result, the value assessment it performs is not only based on past data, but also reflects the current state of the market, helping stakeholders better understand the ever-changing environment. Real-time adaptation distinguishes artificial intelligence from traditional approaches and emphasizes the importance of its application in a rapidly developing market and is a sign of its revolutionary power [3].

We can highlight the following advantages of using artificial intelligence in the process of determining the value of real estate.

High level of accuracy. The use of artificial intelligence opens up the potential to avoid the subjectivity and bias that have long plagued conventional property valuation methodologies. This is one of the most convincing and obvious advantages of this approach. Artificial intelligence algorithms operate in a world of unadulterated facts, free from the influence of subjective opinions or emotional influence.

This objectivity ensures that the value calculation will be based solely on the data itself, resulting in a valuation that extremely accurately reflects the property's actual market value. This gives stakeholders confidence in the honesty and fairness of the deal.

Time efficiency. In the context of property valuation, time is an extremely valuable resource. Traditional methods require days or even weeks to establish a thorough assessment, which has been arguably surpassed in terms of efficiency by AI-driven processes. The time required for assessment is greatly reduced thanks to the unparalleled speed with which AI systems analyze data. This approach not only speeds up transactions, but also adds adaptability to the valuation process, allowing people to quickly respond to changes in market conditions.

Cost reduction. The effectiveness of artificial intelligence also affects the financial component of the property valuation process. Traditional approaches often require multiple evaluations and even repeated re-evaluations to improve accuracy, which can lead to significant costs. Instead, the application of machine learning techniques reduces the need for repeated evaluations by efficiently finding trends from different data sources and synthesizing them. As a result, this leads to a significant reduction in costs for both buyers and sellers, as well as reallocation of funds to more important processes.

Use of geodata. Geospatial data is another vital factor in property valuation, covering everything from a property's proximity to amenities to its location in flood zones or near industrial areas. Artificial intelligence algorithms can integrate geographic information systems to incorporate these factors into real-time assessments, offering a level of detail previously difficult to achieve [4].

Clear understanding. The openness of appraisals provides transparency in real estate transactions that was not available using traditional methods. Algorithms used in AI-based evaluation are not “black boxes” because they offer succinct justifications for their conclusions. A detailed explanation of each evaluation methodology is provided to interested parties. Such openness increases the level of trust and facilitates making informed decisions.

Improved investment selection. An empirical basis for AI-based valuation gives stakeholders a powerful tool to improve investment choices. Investors, buyers and sellers can use the information gained through data-driven valuation to make informed decisions. Valuation accuracy provided by artificial intelligence helps reduce risks when determining the expected return on real estate or analyzing market trends to make strategic choices. This way of making decisions prevents the possibility of overpaying for the purchase or underestimating the value of the real estate when selling, which leads to more successful and justified results [2, 5, 6].

2.2. Analysis of available competitive solutions

The most popular applications for selling, buying and renting real estate on the Ukrainian market, such as DIM.RIA [7] and OLX [8], were studied.

A significant advantage of these services is their popularity in Ukraine, which in turn has led to the fact that such services have a huge amount of real estate data available, which is constantly growing. However, fortunately, this data can be used in the process of developing this service, as it can be purchased at your own expense.

As of 09/10/2023, the OLX service contained the following active ads: apartments for sale - 190,447, houses - 73,942, land plots - 35,147, commercial real estate - 50,939, daily rental housing - 16,145. While DIM. RIA: apartments - 220,317, houses - 31,110, land plots - 4,728, commercial properties - 31,548, total properties for sale - 673,499, total properties for rent - 117,873.

2.2.1. OLX

Having analyzed the OLX service, it was concluded that it cannot currently be called a competitor to the system developed in this work, as it serves only as a platform for posting ads. Therefore, it does not contain any analytical tools, indicators, services, etc. Moreover, it covers too many different areas of buying and selling, while this system is highly specialized and focused on a specific market segment.

2.2.2. DIM.RIA

The advantage of the DIM.RIA service is an intuitive interface, the convenience of searching and filtering ads, the availability of a notification function, customer orientation and the establishment of interaction between real estate sales experts and users, display of various indicators of city neighborhoods on a ten-point scale, display of amenities and distances to them for a specific real estate object.

In addition, it should be possible to determine the approximate value of real estate for free for an authorized user, in case he wants to sell it. The description of such real estate includes a large number of different parameters related to the real estate itself (area, number of rooms, floor, availability of balconies, quality of repair, availability of insulation), and many additional factors (availability of an elevator, parking space, playgrounds, etc.) .

It is not known which of the listed parameters and characteristics are actually taken into account in pricing, and which are used only to compile a detailed description of the real estate in the ad. However, looking at the tool itself, one gets the impression that the service only returns the average value of real estate with similar parameters in a certain area on the map, which is not a sufficiently accurate indicator for a correct real estate assessment.

Also, a useful function for the user is the display of the cost of rent and sale of existing real estate according to the selected parameters in the form of a graph. However, such statistical information will not be enough for a buyer to conduct a superficial assessment of a specific real estate object, so this function is still relevant in the Ukrainian market.

In addition, it is also important to analyze the functionality offered by existing competitive applications abroad. Below are the most popular services, as well as special features that make them stand out from the rest.

2.2.3. Zillow

This service [9] is widespread and relevant mainly in the United States of America and Canada.

The unique features of the service are the availability of a schedule of changes in real estate value over time (data is shown only for past years, without predicting future value), as well as an assessment of the area from the point of view of movement by various means of transport. Also, the advantage of this service is the availability of a free property value estimation function based on the entered characteristics for an unauthorized user. However, when testing this function, an error was received about the insufficient amount of data to determine the cost. The following advantages of the application can be distinguished:

- convenient display of data on a real estate unit;
- availability of a schedule of real estate value changes over time (only value history);
- display of useful data about the area where the real estate is located;
- the expected rental value of the property is displayed.

The only drawback is the lack of a graph predicting changes in real estate value in the future, which would be a logical continuation of the display of historical data.

2.2.4. Realtor

Realtor [10] is widespread and relevant only on the territory of the United States of America. It stands out among others with the ability to display various important data on the map. The main advantage is the display of a map showing the location of educational institutions with the possibility of filtering them by rating, food establishments, public transport traffic, bike paths, noise pollution and regions of possible flooding in case of rising water levels, which is very important information for real estate buyers.

2.2.5. Redfin

Redfin [11] is widespread and relevant only on the territory of the United States of America. The service provides a lot of statistical information in the form of graphs, which can be important when making a decision to buy real estate: statistics on real estate in the center of Chicago for the last 30 days; graph of changes in average real estate value in downtown Chicago over the past 5 years; additional statistics on real estate sales in downtown Chicago for the past 3 months.

We can highlight a number of advantages of this application compared to competitors:

- display of statistical data on real estate in the selected area of the city for the last 30 days;
- availability of a schedule of changes in the average value of real estate over time;
- availability of a graph of the number of sold real estate over time;
- the availability of a schedule of the average number of days required to sell real estate.

2.2.6. PropStream

The PropStream service [12] is distinguished by the largest functionality for data analysis. Obviously, this app is specifically designed for use by investors, brokers, real estate agents, and other experts in the field.

This service offers the following functions: display of analytical reports, search of all possible documents related to property, assessment of real estate value, visualization of prospective areas on the map, display of historical graph of price growth, determination of rental value, display of statistical data of the area, monthly changes in value and changes in rent, rental income valuation, equity valuation, growth rate and many more.

Advantages:

- a specialized application that provides many analytical tools useful for investors and various real estate transaction experts;
- availability of tools for email notifications and newsletters;
- the ability to select an area on the map in which real estate should be searched or analyzed;
- a large database with the history of real estate objects.

Disadvantages:

- complex functionality that requires a lot of time and preparation in order to understand it and be able to fully use it;

- outdated user interface;
- absolutely all functions are paid;
- designed for a narrow range of users.

Of all the services, the most functionally similar to the system developed in this research will be PropStream, as it provides a wide range of data analysis functionality. However, it has a complex user interface, and also, at first glance, it does not cover all the functions that should be implemented in this research.

The main difference of the system developed in this research, compared to the existing solutions, is that the real estate evaluation will be carried out not only by the basic physical characteristics of the real estate (for example, such as the area, number of rooms, floor, year of construction, availability of parking space, etc.) , but also based on location. The closest and most important locations with a radius of 500, 1000 and 3000 meters to this object will be analyzed, which can positively or negatively affect its value.

Shopping and entertainment centers, higher education institutions, schools, hospitals, parks, architectural monuments, food establishments (restaurants, cafes), etc., can be attributed to locations that have a positive effect. The value of real estate will be negatively affected by proximity to such locations as a landfill, a cemetery, a highway or a central highway, a chemical plant, or a railroad.

It is important to determine whether such factors really have a significant impact on the final value of the purchase and rental price, and if so, to what extent. On the basis of this data, it is possible to make a more accurate assessment of the value, identify promising areas and predict the rise or fall of the price based on the future development plans of the city.

In order to implement this in practice, it is necessary to collect a large amount of data, starting from the real estate ads themselves for an adequate assessment of the market, ending with statistical data from different cities or districts and data related to important locations. And during the training of the machine learning model experimentally, it is necessary to determine which factors should be taken into account in the evaluations and what their influence is on the overall result.

As a result of the study, it was established that the available applications on the Ukrainian market do not satisfy all the needs of users, as they do not have sufficient functionality for recommendations and analytics. And that is why it is urgent to develop a new service that will take into account these shortcomings and enable users to more effectively make decisions regarding the purchase and sale of real estate.

3. Methods of solving the problem of real estate pricing analysis

3.1. Formation and justification of the problem

The information system will be used to optimize the process of searching for real estate, evaluate its value, and provide the opportunity to make mutually beneficial deals between buyers and sellers. The most important goal of the project is to apply machine learning algorithms to perform the value estimation function, which will be relevant for all real estate market participants. At the same time, it is also important to achieve the highest possible level of forecasting accuracy.

The difficulty of the task lies in the limited number of resources, and especially in the secrecy of real estate data. The disadvantages of the Ukrainian real estate market are that all real estate data is hidden, including historical data. The State Statistics Service of Ukraine provides a limited amount of information in this field, which cannot be used within the scope of this work. For example: index of growth of prices for residential real estate in Ukraine, compared to the previous year; the average cost of renting a one-room apartment in different regions of Ukraine, etc.

In the process of analyzing competitive applications, an example of displaying statistical data related to the city and its districts, as well as historical data of a specific real estate object, was given. This information would be impossible to display without access to historical data that is open, for example, in the United States of America.

Within the framework of the study, it is planned to achieve the accuracy of determining the real estate value, which will be 5-7% on average. The error value will be calculated based on the test data. Ensuring this level of accuracy of predictions will be considered sufficient, since the costs of realtor services are on average 3-5% of the total cost of real estate sales [13], and therefore are almost completely covered by the error in calculations.

Competing applications abroad offer slightly better price prediction results, which is achieved due to the openness of the data. For example, Zillow's Zestimate tool guarantees such accurate predictions that the median error value will be only 3.2%. Although in case of insufficient amount of information in certain regions, this value can increase up to 7.52% [14].

However, given the fact that today there are no services with similar functionality available for use on the territory of Ukraine, the current predicted level of error for the system under development is considered acceptable.

3.2. Building models for problem solving

The process of creating a machine learning model is conditionally divided into several stages, while the correct and successful execution of each previous stage has a significant impact on the next one, as well as on the final result in general.

The first stage. First you need to find all available data sources that can be used in the work. Such data should be as much as possible, since the amount is an important criterion when building a machine learning model. In addition, it is necessary to assess the quality and completeness of the data of each source. In this case, the advertisement for the sale of real estate must contain such parameters as the total area, kitchen area, floor, total number of floors of the house, year of construction of the house, address of the property, etc.

Once you've found a quality data source, you need to upload that data locally. For this, you should use the official API of the source. In this case, such services are not free, and therefore you also need to pay for a key that will provide access to the necessary information.

After that, you need to filter the data: find and remove duplicate ads, as well as identify outliers - such data objects that stand out too much from the rest. The simplest outlier is a property whose price per square meter is several times higher than the vast majority of records, while there are no other data with similar values in the sample.

Next, you need to check if there are any missing data in the ads and which ones. If not too important data is omitted, they can be filled in, otherwise such records must be filtered as well. The process of filling in missing data will be described in the practical part of model creation.

Also, for each real estate object, you need to unify the data, i.e. convert them to the form used in the system. The next step is to determine the coordinates of the property at the address using an external API. Finally, you need to add the geographic data of the property, such as the number of bus stops nearby, the number of hospitals, schools, etc.

The second stage. The next stage involves determining a minimum list of the most important features and parameters of real estate that affect the machine learning model's ability to predict real estate value. Usually, this procedure is carried out empirically, since it is almost impossible to predict how a machine learning model will behave when operating on a specific set of data.

Some individual characteristics of real estate need to be modified in such a way that they are processed more effectively by the model. In this case, it would be advisable to apply the method of One-Hot Encoding, which involves the representation of complex categorical features in the form of simple values 0 or 1. For example, instead of listing all types of walls of the house in one column (brick, gas block, reinforced concrete, foam block, etc.), you can create a separate column for each type of wall, for example "brick_wall", the value of which will be 0 or 1. At the same time, if the number of such columns will be large, you can try to group them by common features. This approach can have a positive effect on the process of model training.

The third stage. Here you need to divide the data into two samples - training and testing, usually this is done in a percentage ratio of 80%-20%, respectively. It is desirable that the data be chaotic, that is, not sorted and not grouped by any of the parameters. This especially applies to those parameters used in the process of training the model.

In order to achieve the best results, you can try to select data for testing by the iterative method, iterating through all possible options. This method is called cross-validation. The method consists in taking 20% of the test data from the total hundred in turn and training the model with the same algorithm on the same data, but choosing different data for testing: 0-20%, 20-40%, 40-60%, 60 -80%, 80-100%. This way you can make sure that the best data is selected for training and also prevent overtraining of the model.

The very process of training machine learning models can be attributed to the same stage. Considering that the number of active ads for one large city ranges from several thousand to several tens of thousands (before filtering), this number is considered relatively small and the expected training time of the model should last seconds or minutes. Therefore, the optimal solution would be not to limit yourself to one specific algorithm, but to try several at once and check which of them will work more efficiently. At this stage, the algorithms should be run with the parameters set by default.

The fourth stage. Now it is necessary to evaluate the accuracy of the models using one of the quality metrics that will be described in the subsection "Selection and justification of problem solving methods". After learning which algorithms work best for solving this task, you should choose one or more of them and try to find the optimal hyperparameters. In this way, it is possible to achieve high accuracy of the model for a specific task.

It is obvious that the regression algorithms of machine learning with this approach will be taken from ready-made libraries, since the specified sequence of steps involves simultaneous work with several of them.

It was decided to limit the number of cities for calculating the value of real estate to one. This is due to many good reasons, the most important of which are:

- the absence of free data in public access, which means the inevitability of attracting own funds for the development of the system;
- the need to build a separate machine learning model for each city, since the cost of real estate varies greatly for different cities.
- the process of creating machine learning models is universal, so creating new models for other cities is just a matter of time and attracting additional funds and computing resources.

In addition, the value of the work is to make it possible to determine the value of real estate by means of artificial intelligence.

3.3. Selection and justification of problem solving methods

Since this work involves determining the value of real estate based on certain input parameters, regression algorithms help to solve this type of task [15-28].

Regression – an approach that involves investigating the relationship between independent variables or traits and a dependent variable or outcome. Regression is used in machine learning as a method of predictive modeling, that is, to predict or predict outcomes. Regression is used to solve many different machine learning tasks in various fields. Regression analysis is used for financial forecasting, forecasting trends in healthcare, manufacturing, commerce, entertainment, sports, etc. For example, it is used in various systems to predict housing prices, stock values, and changes in wages.

Most often, regression models of machine learning are used in:

- forecasting constantly changing values, such as housing prices, stock prices, wage fluctuations;
- predicting the success of future retail sales or marketing campaigns, in order to ensure effective use of resources;
- predicting customer or user behavior trends, for example in streaming services or e-commerce;
- analysis of datasets to establish relationships between variables and the result;
- forecasting interest rates or share prices based on various factors;

- creation of visualizations of time series.

Algorithms work in such a way that, using machine learning models, they "understand" the relationship between independent variables and the result (or dependent variable). The ready-made model can be used to predict the results of new, not involved in the learning process of input data or in the case when the data is incomplete and you need to fill the gaps in them.

Since regression involves training "with a teacher", the creation of machine learning models requires the availability of labeled input and output training data. Such data are extremely important and play a crucial role in the process of training models and, accordingly, in predicting results.

The amount of training data should be large enough to understand how the parameters depend on each other, but at the same time not too large, as this can lead to "overtraining" of the model, which will deprive it of flexibility. In addition to quantity, training data must also be of quality. They must be complete, consistent, current, relevant, etc. Even a small part of low-quality training data can spoil the entire sample and lead to the fact that the machine learning model trained on it will predict the data incorrectly [15].

3.3.1. Types of regression

In machine learning, there are a number of different approaches used to perform regression, and there are also different popular algorithms. Different techniques may include different numbers of independent variables or handle different types of data. Different types of machine learning regression models can predict different relationships between independent and dependent variables. For example, linear regression techniques assume that the relationship is linear, so this approach will not be effective for data sets with non-linear relationships.

Linear Regression — is a statistical method of establishing a relationship between two variables using a straight line. The line is drawn by finding the slope and intercept that defines the line and minimizes the regression errors.

The simplest form of simple linear regression has only one variable x and one variable y. The variable “x” is the independent variable because it does not depend on what you are trying to predict with the dependent variable. The variable “y” is a dependent variable because it depends on what you are trying to predict [15-19].

Multiple linear regression (MLR). When predicting the outcome of a complex process, it is best to use multiple linear regression instead of simple regression. At the same time, it is not necessary to use complex algorithms to perform simple tasks.

A simple linear regression can accurately capture the relationship between two variables in simple relationships. But when you are dealing with complex problems, where several independent parameters affect the result, then you need to go from simple to multiple regression.

The multiple regression model uses a formula containing more than one independent variable, thus, it is able to work with curves, as well as with non-linear dependencies [20].

Model building in linear regression occurs by minimizing the sum of squared deviations between observed and estimated values (the method of least squares).

Decision tree. This is a tree, the leaves of which contain the values of the objective function, and the nodes contain transition conditions (for example, "GENDER is MALE"), which determine which of the edges to move along. If the condition is true for a given observation, then the transition is made along the left edge, if it is false, then along the right edge.

The decision tree grows by iteratively splitting its nodes until the "leaves" contain no more branches (the answer to the last question gives an unambiguous result), or until some termination condition is met. The creation of a decision tree starts from the root of the tree, and the distribution of data takes place in such a way as to obtain the largest value of information gain (IG) [19]. A data set is pure or homogeneous if it contains only one class (YES or NO). If the data set contains several classes, the table is impure or heterogeneous (a combination of YES and NO) [23,24].

Random Forest. By combining several uncorrelated decision trees, a significant increase in model accuracy can often be achieved. This method is called random forest. During growth, the branching of the tree is affected by certain random processes (this is called randomization). The final model reflects tree averaging.

There are different methods of randomization. According to Breiman, who coined the term "random forest" in 1999. The process of creating a random forest is as follows. First, a random instance is selected from the total data set for each tree. As the tree grows, a subset of special features is selected at each node. This serves as a criterion for dividing the data set. After that, a target value is determined separately for each decision tree. Averaging the values of these predictions represents the final prediction of the random forest algorithm [19].

Since this algorithm combines several models into one, it belongs to the field of "ensemble learning". And to be more precise, Random Forest is the so-called bagging technique.

In addition to Bagging, the most famous type of ensemble learning technique is Boosting, and the most famous algorithms in this area are the AdaBoost and XGboost algorithms.

In order to better understand the essence of the Random Forest algorithm, it is worth investigating the concept of "ensemble learning" in more detail, as well as consider its varieties.

An ensemble is a collection of prediction tools that give an answer together (for example, the average of all predictions can be taken as a result). The reason why ensembles should be used is simple - multiple prediction tools trying to obtain the same variable will give a more accurate result than one. Ensemble techniques are further divided into bagging and boosting.

Bagging is a simple technique that constructs independent models and combines them using some averaging model (eg, weighted average, majority voting, or normal average). Usually, a random sample of data is taken for each model, since all models are slightly different from each other. The sample is constructed according to the return selection model. Because this technique uses multiple uncorrelated models to build a final model, it reduces variance.

Boosting is an ensemble building technique in which prediction tools are built sequentially rather than independently. This technique uses the idea that the next model will learn from the mistakes of the previous one. They have an unequal opportunity for errors to appear in subsequent models, and those that give the largest error will appear more often. Prediction tools can be chosen from a wide range of models, such as decision trees, regression, classifiers, etc. Because the forecasters learn from the mistakes made by the previous ones, it takes less time to get to the real answer. But you need to choose the stopping criterion carefully, otherwise it can lead to overtraining. An example of boosting is gradient boosting [29, 30].

Support Vector Regression. The functionality of the support vector regression method (SVR) is based on the Support Vector Machine (SVM). Let's consider a simple example. You need to find a linear function:

$$f(x) = \langle \mathrm{w}, \mathrm{x} \rangle + \mathrm{b}\ , \tag{1}$$

where $\langle \mathrm{w}, \mathrm{x} \rangle$ describes the cross product.

The goal of SVR is to find a straight line as a model for the data points, while the line parameters should be determined so that the line is as "flat" as possible. This can be achieved by minimizing the norm:

$$\|w\|_2 \coloneqq \sqrt{(w_1)^2 + (w_2)^2 + \cdots + (w_n)^2} = (\textstyle\sum_{i=1}^{n} (w_i)^2)^{1/2}. \tag{2}$$

In the model building process, it does not matter how far the data points are from the modeled line, as long as they are within a defined range (from -ε to +ε). Deviations exceeding the specified limit **ε** are not allowed [19].

K-Nearest Neighbors (KNN). KNN is a non-parametric, simple but powerful supervised learning algorithm that can be used for both regression and classification tasks. The basic idea of KNN is to find the K closest data points in the training space for a new data point. Having done this, it is possible to assign a new data point to one of the classes by analyzing which class prevails among k nearest neighboring points [31]. In the case of regression, the dependent variable is continuous, it is scattered over the entire coordinate plane. When there is a new data point, the number of neighbors (K) is determined using any distance metric. After finding neighbors, the predicted value of the new data point is the average of all compatible neighbors.

For example, consider the house price forecast. The price is the dependent variable and the area in square meters of the house is the independent variable. Now after plotting all the data points on the Cartesian plane, when a new square meter point appears, the average value of the K neighbors per square meter of the house is the value of the new data point. Therefore, instead of predicting the class, the regressor uses the average value of all neighboring points [31].

Selection of K value. The value of K is the main part of KNN and its selection can be difficult. The sequence of finding the best and optimal value of K:

- it is necessary to divide the data set into training and testing samples;
- select a range of K values. You can start with K = 1 and gradually increase it;
- start training the KNN model for each value of K;
- evaluate the performance of models that have been trained using a range of K values.

It is important to note that the choice of the value of K depends on the data set and the problem itself. A small value of K can lead to the fact that the model is not flexible enough, which characterizes the concept of overfitting, while a large value can lead to underfitting. Therefore, it is recommended to experiment with different values of k to find the optimal one for a particular data set.

Advantages of the algorithm:

- Flexibility: KNN can be used for both regression and classification tasks.
- No need for training: which saves time and computing resources.
- Robustness to noisy data: as it relies on the majority of nearest neighbor votes. This makes it less vulnerable to outliers.
- Works well with small data: No large amount of data is required for prediction. Non-parametric: KNN is a non-parametric algorithm, meaning it does not require any assumptions about the data.
- Simplicity: KNN is a simple and straightforward algorithm.

Disadvantages:

- Optimal value of K: Choosing the right value of K is important because it will affect the performance of the model. There is no universal value for K, and this value depends on the characteristics of the data set elements.
- Imbalanced data: KNN may have bias in case of unbalanced data. That is, when one class should have more examples than another, KNN can predict the majority class for the test examples.

Dimensionality problem: KNN can suffer from high dimensionality problem which occurs when the number of features is large. As the number of measurements increases, the distance between any two points in the data tends to increase, making it difficult to find meaningful nearest neighbors [31].

3.3.2. Quality metrics of regression models

In order for the linear regression model to be applied in practice, it is first necessary to evaluate its quality. For this, there are a number of indicators, each of which is intended for use in different situations and has its own application features (linear and non-linear regressions, resistant to anomalies, absolute and relative, etc.). The correct choice of measure for the purpose of evaluating the quality of the model is one of the most important success factors in the process of solving data analysis tasks.

A "good" analytical model must satisfy two requirements, which often conflict with each other - to fit the data as well as possible and to be convenient for user interpretation. Increasing the fit of the model to the data is usually associated with its complication, since in the case of regression this means an increase in the number of input variables of the model. And the more complex the model, the more difficult it is to interpret.

Therefore, when choosing between a simple and a complex model, the latter should significantly increase the fit of the model to the data in order to justify the increase in complexity and the corresponding decrease in interpretability. If this condition is not met, a simpler model should be chosen [32].

Thus, in order to assess how much increasing the complexity of the model increases its accuracy, it is necessary to use the tool for assessing the quality of regression models (Table 1).

Table 1
Comparison of metrics

Metrics	Advantages	Disadvantages
Mean Squared Error (MSE)	Emphasizes large deviations, and provides a simple calculation.	Tends to underestimate model quality sensitive to outliers. Complexity of interpretation due to quadratic dependence.
Root Mean Squared Error (RMSE)	Ease of interpretation because it is measured in the same units as the target variable.	Tends to underestimate model quality sensitive to outliers.
Mean Squared Percentage Error (MSPE)	Insensitive to outliers. It is well interpreted because it has a linear character.	Since the contribution of all errors of individual observations is weighted equally, it does not allow to highlight large and small errors.
Mean Absolute Percentage Error (MAPE)	It is a dimensionless quantity, so its interpretation depends on the subject area.	Cannot be used for observations where the output variable values are zero.
Symmetric Mean Absolute Percentage Error (SMAPE)	Allows you to correctly work with predicted values, regardless of whether they are greater than the actual value or less.	Approaching zero of the actual or predicted value leads to a sharp increase in error, since the denominator contains both the actual and the predicted value.
Mean absolute scaled error (MASE)	It does not depend on the scale of the data and is also symmetrical: positive and negative deviations from the actual value are taken into account equally. Resistant to outliers. Makes it possible to compare models.	Complexity of interpretation.
Mean Relative Error (MRE)	Provides an estimate of the magnitude of the error relative to the value of the target variable.	Not applicable for observations with a zero value of the output variable.
Root Mean Squared Logarithmic Error (RMSLE)	Logarithmization helps make the magnitude of the error more robust when the difference between the actual and predicted values differs by an order of magnitude or more.	Interpretation may be difficult due to non-linearity.
Coefficient of determination R-squared	Universality, ease of interpretation.	It grows even when additional variables are included in the model. It works poorly when the input variables are dependent.
Adjasted coefficient of determinftion	Correctly reflects the contribution of each variable to the model.	Works poorly when input variables are dependent.

3.4. Development of problem solving algorithms

Some sample data require filling in missing values to avoid filtering. This can be done using the mean, mode, and median values of the total sample [33].

Mean is the arithmetic mean of a set of data. It should be used when the sample has a normal data distribution and when it has no outliers and abnormal values.

Median is the value in the middle of a list of numbers when the data set is sorted in ascending order. It is used when the distribution is not normal and the sample contains outliers.

The mode is the value that appears the most times in the data set. It is relevant when analyzing categorical data to determine the category that occurs most often.

So, from the definition of these concepts and recommendations for use, it can be concluded that the choice of the optimal approach of averaging values depends on the type of data, as well as on their distribution in the sample.

In this work, only one metric for assessing the quality of the machine learning model will be used, namely R-squared. This metric is also called the coefficient of determination, which shows the proportion of the variance of the dependent variable that is explained by the regression model. The peculiarity of this approach is that it shows the value of how much this model works better than the model in which there is only a constant, and the input variables are absent or their regression coefficients are equal to zero.

The most general formula for calculating the coefficient of determination is [34]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2}{\sum_{i=1}^{n}(\overline{y_i} - y_i)^2}\ , \tag{3}$$

where y_i – is the actual value; $\widehat{y_i}$ – the value that was calculated by the model; $\overline{y_i}$ – is calculated according to the arithmetic mean formula:

$$\overline{y_i} = \frac{1}{n}\sum_{i=1}^{n} y_i\ . \tag{4}$$

Based on formula (3), in the context of creating a machine learning regression model, a simplified definition of the coefficient of determination can be given - it is a metric that shows how much the values predicted by the regressor are "better" than if the model simply calculated the average value over the entire sample.

To evaluate the result returned by the R-squared formula, the following scale is used in practice [34]:

- values less than 0.5 (including negative numbers) indicate that the model is bad;
- values >0.5 indicate the satisfaction of the model;
- if the coefficient of determination is >0.8, the model is considered good.

It can be understood from the formula that when the predicted values of $\widehat{y_i}$ are equal to the arithmetic mean $\overline{y_i}$, the result of calculations in the fraction will be equal to 1, and therefore the coefficient of determination will be equal to 0, which indicates poor performance of the model. On the other hand, if $\widehat{y_i}$ are equal to y_i, then this will give 0 in the numerator and the coefficient of determination will be equal to 1, which will mean that the model perfectly predicts all values [34].

4. Testing of the developed real estate value forecasting algorithm

The testing process is important because it provides an opportunity to identify and correct errors in the system at the stage of its development, and also helps to understand how ready the system is for use.

4.1. Search, filtering and data processing

The first and extremely important step, which determines all further possibilities of the system, is the search for up-to-date real estate market data. The amount and quality of data available for training and testing a machine learning model play a significant role in its effectiveness. Data can take different forms, such as numerical, categorical (such as can be used to classify objects based on their belonging to categories) or time series data (a sequence of values collected according to a certain time interval), and can come from from different sources, such as databases, spreadsheets or API [35].

It is necessary to find as many sources of information as possible, which will ensure the availability of a sufficient amount of data for the formation of a high-quality model of machine learning. At the same time, it is also desirable to update this data with a certain periodicity, in order to have further opportunities for their deeper analysis, to provide the notification function and, in general, to retrain the model on current data.

As part of this work, it was decided to collect all the necessary data for only one specific city, because the data collection process is very resource-intensive, and especially because a separate machine learning model needs to be trained for each city, in order to ensure more accurate operation of the program. This is due to the fact that the price of real estate is very different in different cities of Ukraine. For example, the average cost of a one-room apartment in the Lviv region is $61,616, while in the Mykolaiv region it is $23,144, according to open sources of information as of August 2023 [36]. Therefore, in the case of creating one universal model for the whole of Ukraine, it will not be able to conduct an adequate assessment of the cost based on the input data, and it will also lead to unpredictable results for model outliers.

Several different data collection approaches were used. The first of them is data extraction using the official APIs of the following websites: DIM.RIA [7] and OLX [8]. The following Telegram channels were also found that publish real estate sales and rental ads: Real Estate Kyiv and Oblast [37], Real Estate Kyiv Oblast [38].

With the help of the appropriate Telethon library [39], it was possible to collect and process these data for further use. A code snippet that analyzes the text of a real estate rental ad and extracts data from it:

```
message_text = message_text.replace('\n', ' ')
address_pattern = r'\[.+\]'
square_pattern = r'[0-9 ]+м²'
ad_id_pattern = r'№[0-9a-zA-Z ]+'
price_pattern = r'Ціна[0-9a-zA-Z \t]+'
floor_pattern = r'[0-9й \t]+поверх'
rooms_pattern = r'#[0-9]кімнатна'

address = re.search(address_pattern, message_text).group(0) if re.search(address_pattern,
message_text) else 'None'
square = re.search(square_pattern, message_text).group(0) if re.search(square_pattern,
message_text) else 'None'
square = square.replace('м²', '').strip()
ad_id = re.search(ad_id_pattern, message_text).group(0) if re.search(ad_id_pattern,
message_text) else 'None'
price = re.search(price_pattern, message_text).group(0) if re.search(price_pattern,
message_text) else 'None'
floor = re.search(floor_pattern, message_text).group(0) if re.search(floor_pattern, message_text)
else 'None'
rooms = re.search(rooms_pattern, message_text).group(0) if re.search(rooms_pattern,
message_text) else 'None'

return ad_id, square, price, address, floor, rooms, message_text
```

After data collection and processing, it is also necessary to filter them. Some data are of no practical value because they contain insufficient information in the description of the real estate unit. Therefore, if one of the above fields was not filled in, such data were not taken into account.

However, there are exceptions where data with missing values can be used in the model building process.

In the event that the value of kitchen area or living area is missing in a real estate unit, such data were filled with the median of these values in the total sample. It was inappropriate to take the arithmetic mean value, since the distribution of these data is skewed.

In addition, the heating type, ad type, and wall type values have been filled in by the mode of the data in the sample, since these are categorical data types. The year of construction of the house was also filled in with the help of mode:

```
df.loc[df['kitchen_square']!= -1, 'kitchen_square'].median()
df.loc[(df['living_square']!= -1), 'living_square'].median()
df.loc[(df['wall_type']!= 'None'), 'wall_type'].mode()
df.loc[(df['building_year']!= -1), 'building_year'].mode()
df.loc[(df['advertisement_type']!= -1), 'advertisement_type'].mode()
df.loc[(df['heating_type']!= -1), 'heating_type'].mode()
```

In this way, the following data were determined to supplement information about real estate in the city of Kyiv: kitchen area - 13 square meters, living area - 30, type of walls - red brick, year of construction of the house - 2013, type of advertisement - from an intermediary, type of heating - centralized.

Open source OpenStreetMap Nominatim API [40] was used to translate the real estate address into coordinates on the map. In order to use it, no registration is required, just send a GET request and add the address to the link:

```
https://nominatim.openstreetmap.org/search?q=Хрещатик+Київ+18&format=geojson
```

In this way, it was determined that the coordinates of the address of Khreshchatyk Street, 18, in the city of Kyiv are: latitude - 50.4510346, longitude - 30.523997621368913.

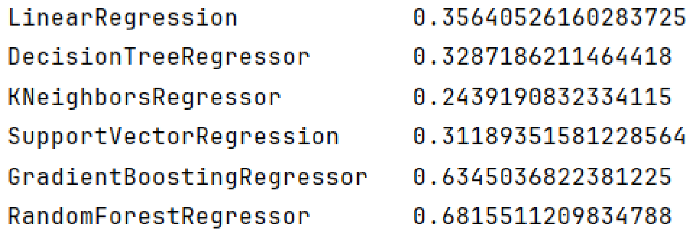
The next step is to find all important locations within a given radius near the property. For this, an open source of data was used - Python Overpass API [41]. To use this API, you need to create a corresponding request containing the coordinates of a specific point, as well as the radius within which you need to search for the nearest locations. Below is an example of searching for locations on the map within a radius of 100 meters to a given point:

```
lat = 50.450310
lon = 30.523736
query = """
   (node(around:100,{lat},{lon});
   );out;
   """.format(lat=lat,lon=lon)
api = overpy.Overpass()
result = api.query(query)
file_path = "overpass_results.txt"
with open(file_path, "w", encoding="utf-8") as file:
   filtered_nodes = [node for node in result.nodes if node.tags]
   for node in filtered_nodes:
      file.write(f"Node {node.id} at ({node.lat}), {node.lon}, tags: {node.tags})\n")
```

127 locations within a radius of 100 meters of this address were received. Obviously, this data will need to be filtered, since the list of locations includes trees, garbage cans, street lamps, mailboxes, benches, etc. However, among them there are also important locations that should influence the value of real estate.

4.2. Training a machine learning model

Now that the data has been collected, filtered and supplemented with additional parameters, you should proceed directly to training the model. It was decided to use several machine learning regression algorithms at once, with the aim of finding the one that best copes with solving this problem. Algorithms were run with standard parameters, the result is presented on Figure 1.



```
LinearRegression            0.35640526160283725
DecisionTreeRegressor       0.3287186211464418
KNeighborsRegressor         0.2439190832334115
SupportVectorRegression     0.31189351581228564
GradientBoostingRegressor   0.6345036822381225
RandomForestRegressor       0.6815511209834788
```

Figure 1: Comparison of the accuracy of models trained by different algorithms using the R-squared metric

Comparing the coefficients of determination of the obtained results, it became clear that the "random forest" algorithm showed the best result among the rest. Therefore, further development of the model and selection of hyperparameters will be performed specifically for this algorithm.

The following hyperparameters were empirically selected, which showed improved results in predicting predicted values using the random forest algorithm:

```
predictor_rf = RandomForestRegressor(n_estimators=200,
                                     min_samples_leaf=2,
                                     max_depth=6)
```

Due to the configuration of hyperparameters, as well as using the method of cross-validation of data, the value of R-square could be increased to 0.81, which is considered a high indicator.

To test the machine learning model, a validation sample of real estate data was created and tested to see how the model performs in predicting value compared to real data. On Figure 2 the vertical axis represents the real estate value per square meter in US dollars, and the horizontal axis represents the proportion of the number of advertisements.

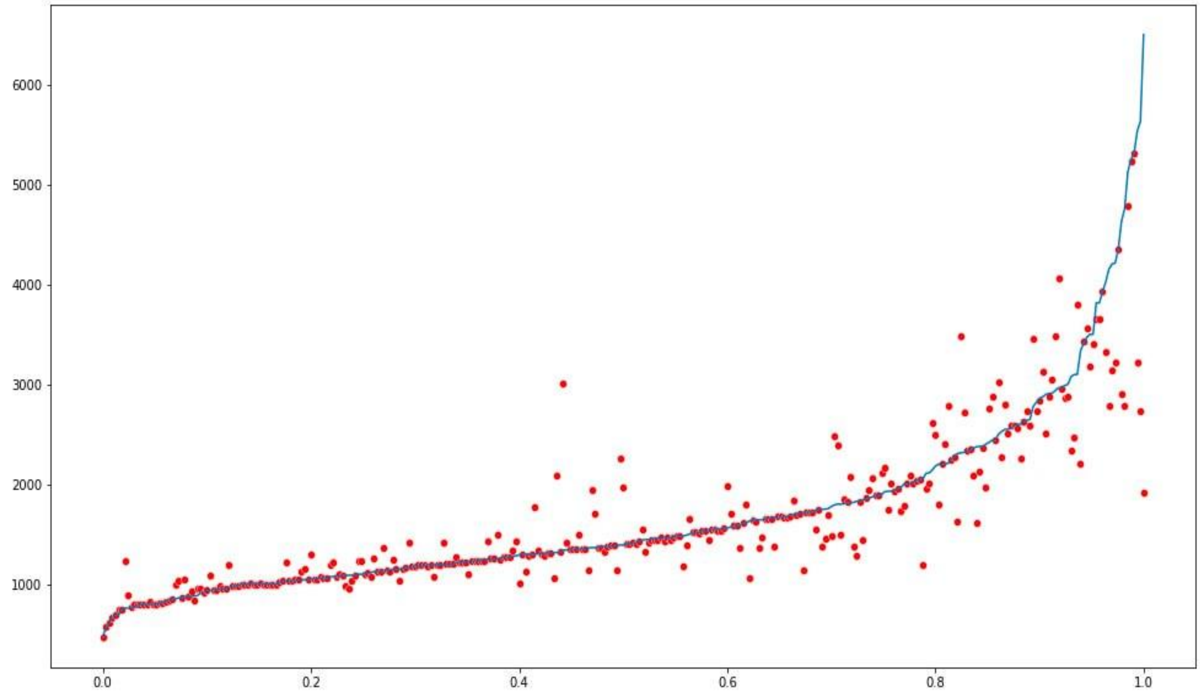


Figure 2: The result of forecasting the value of a square meter of real estate according to its parameters in Kyiv by the developed model

The red points show the value values predicted by the model, and the blue ones show the real values.

The graph shows a phenomenon called heteroskedasticity, which means that as the dependent variable increases, the predicted variable will have larger deviations from the actual values. That is, with an increase in the value of real estate per square meter, the variability of the predicted values increases. This problem is related to the insufficient amount of data that was used when training the model, and this is a normal phenomenon. Most of the properties in the data sample have a price per square meter of less than $2,000, so it is difficult for the model to predict such unique cases. In the future, this problem can be solved by collecting more and more data about real estate for sale.

Figure 3 shows the distribution of the absolute error of the predicted results in percentages. The green line represents the mean error value, and the red line represents the median.

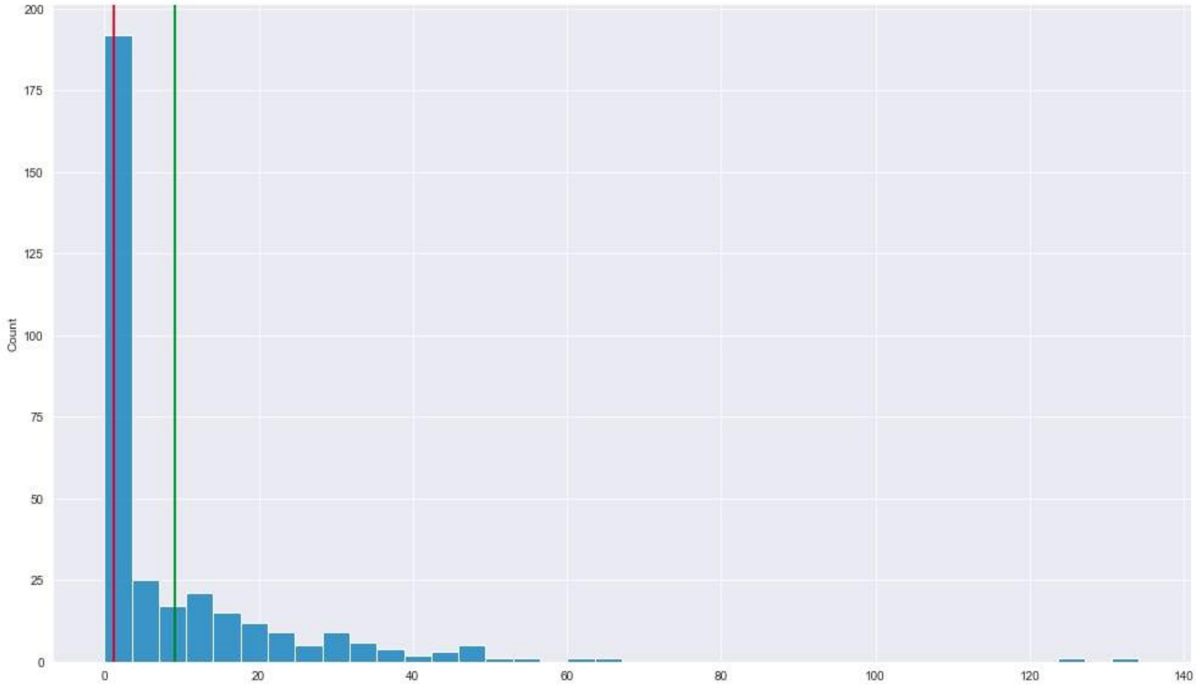


Figure 3: Distribution of the absolute error of forecasting the value of a square meter of real estate according to its parameters in Kyiv

In order to better evaluate the results of the calculation of the absolute error, the data are presented in the following way:

APE > 50% for 0.012084592145015106 of test data
APE > 20% for 0.1933534743202417 of test data
APE > 10% for 0.31722054380664655 of test data
APE > 5% for 0.4108761329305136 of test data
APE > 1% for 0.5347432024169184 of test data
APE < 1% for 0.4652567975830816 of test data

Where APE is the absolute error calculated by the formula:

$$APE = \left|\frac{d_i - \widehat{d_\imath}}{d_i}\right| , \quad (5)$$

where d_i – a real value; $\widehat{d_\imath}$ – the value calculated by the model..

Now the percentage of absolute error of forecasting the real estate value can be represented in the form of a diagram (Figure 4).

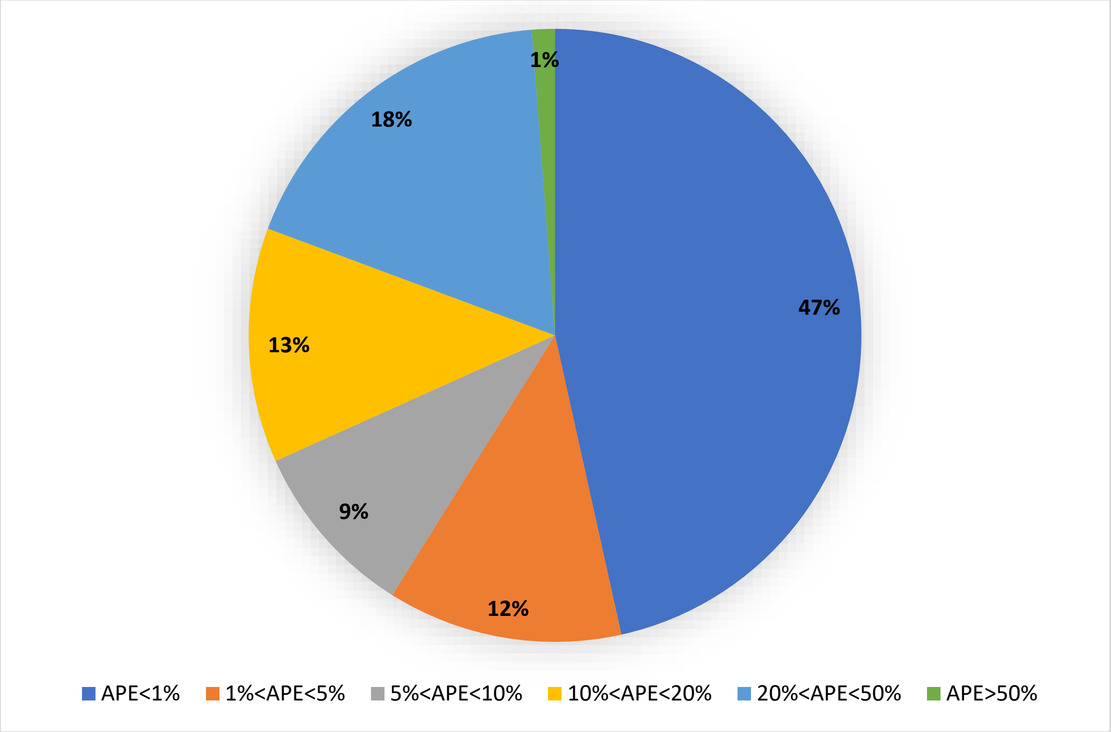


Figure 4: The absolute error of the predicted real estate value

Summarizing all the information related to the machine learning model, we can say that in general the model performed well, the results of the study indicate that it was successful. On condition of constant storage of new data, as well as retraining of the model on them, one can even hope for better results of its work. Using the same approach, it is possible to develop an effective model for any other city of Ukraine, provided there is a sufficient amount of data.

5. The Architecture of the System Prototype

Based on the tasks and functions that the program should perform, as well as the intended users and their capabilities, a diagram of use cases was built (Figure 5). This diagram serves to visualize the user and functional requirements of the system.

The Figure 5 shows four actors of the system: unauthorized user, user (authorized), investor, administrator. Each of them can interact with the system in different ways and has its own set of capabilities and functions. If necessary, the user of the system can increase his role in the hierarchy to expand the number of functions available to him. Below is a more detailed description of the actors of the system:

Unauthorized user – has the ability to view real estate for sale in a grid and on a map and use filtering tools by parameters. You can also view complete information about the real estate unit, including the contact details of the ad author.

User (or authorized user) – has the same functions as an unauthorized user, but has much more options. He can create, edit and delete his own real estate ads. In the case of creating an ad, the system will indicate the expected value of the property according to the entered parameters and the most important factors that influenced the estimate. Such a user can add other people's ads to the list of favorites, view it and manage it. In addition, the possibility to display recommended real estate according to a given filter has been introduced.

Investor – a unique system role that extends the role of an authorized user by providing access to analytics tools. Can view graphs, determine the cost of renting a real estate unit and calculate the payback period through rent. Has the ability to connect notifications about the appearance of recommended real estate according to the specified filters.

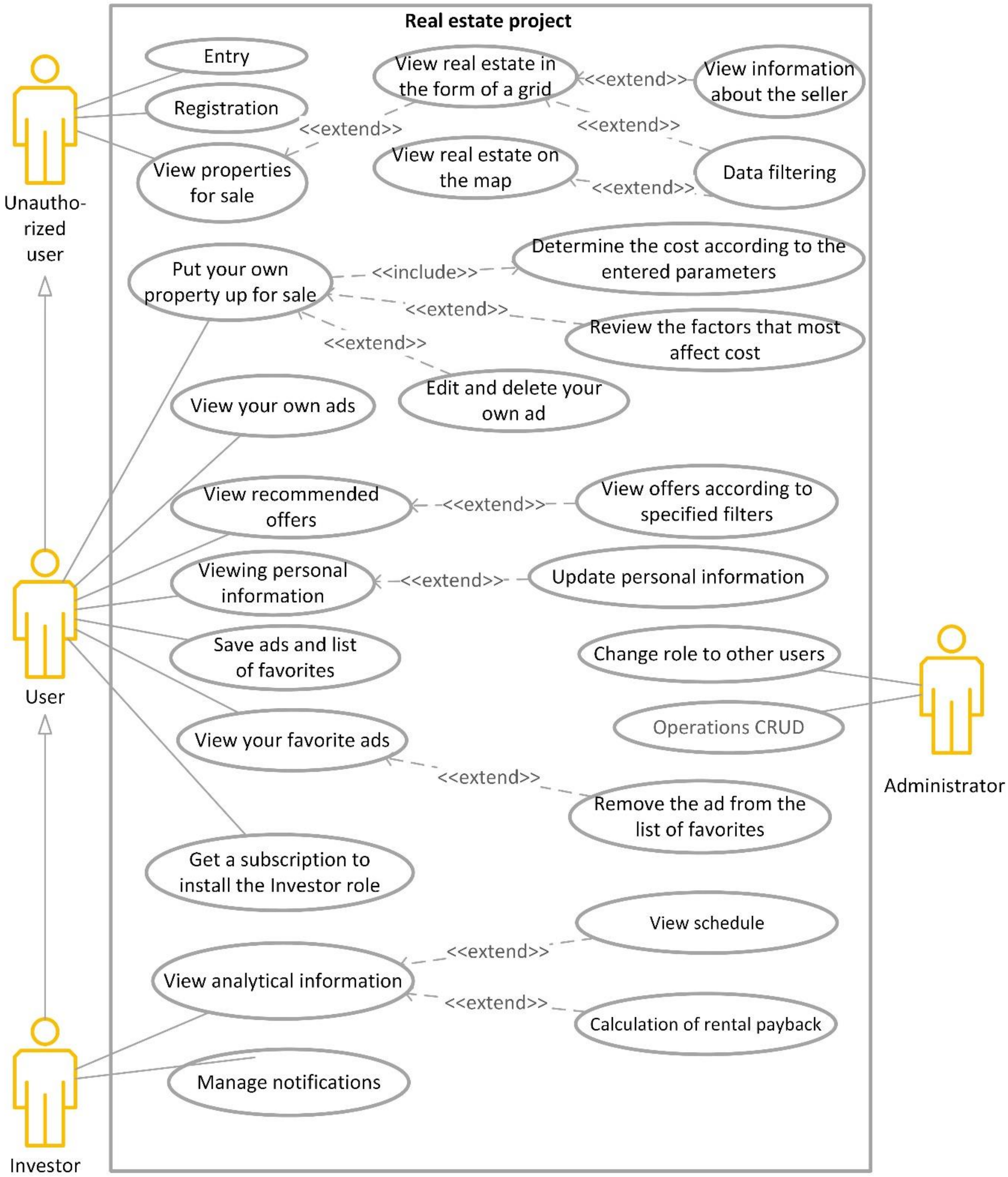

Real estate project
Entry
Registration
View properties for sale
View real estate in the form of a grid
View real estate on the map
View information about the seller
Data filtering
<<extend>>
<<extend>>
<<extend>>
<<extend>>
Unauthorized user
Put your own property up for sale
<<include>>
Determine the cost according to the entered parameters
<<extend>>
Review the factors that most affect cost
<<extend>>
Edit and delete your own ad
View your own ads
View recommended offers
<<extend>>
View offers according to specified filters
Viewing personal information
<<extend>>
Update personal information
Save ads and list of favorites
Change role to other users
Operations CRUD
Administrator
User
View your favorite ads
<<extend>>
Remove the ad from the list of favorites
Get a subscription to install the Investor role
View schedule
<<extend>>
View analytical information
<<extend>>
Calculation of rental payback
Manage notifications
Investor

Figure 5: Use case diagram

Administrator – is responsible for managing all system entities, including the ability to change user roles. He is the person who maintains the system and helps users with editing parameters in case of errors.

To implement this system, it was decided to use a microservice architecture, which will ensure simplicity, flexibility and scalability [42-45]. The software product will consist of three separate independent services connected to each other.

Figure 6 shows the general scheme of the system, its components and the interaction between them. Each service is discussed in detail below to gain a deeper understanding of their tasks and functions.

Data providing service. This service is responsible for searching, filtering and initial data processing.

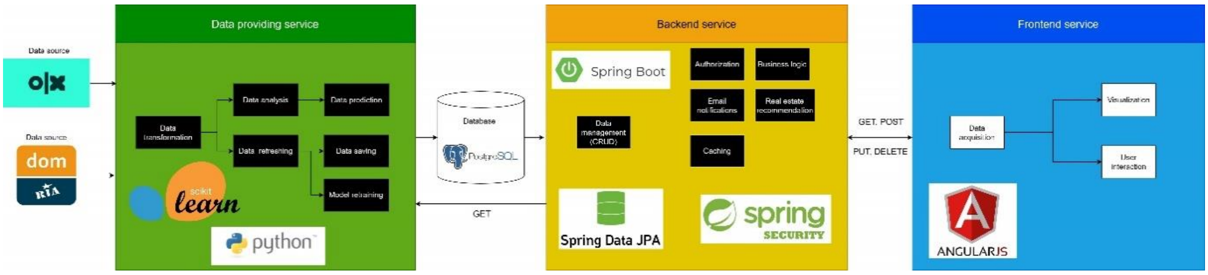


Figure 6: General structure of the system

Also, its task is to constantly update data with a certain frequency, in order to monitor the current state of the market, as well as provide the application with up-to-date data, which is especially important for ads. This data is saved to the database and is available to other parts of the application. In addition, the service analyzes the data, builds machine learning models based on them and predicts the value of real estate based on the parameters given to it.

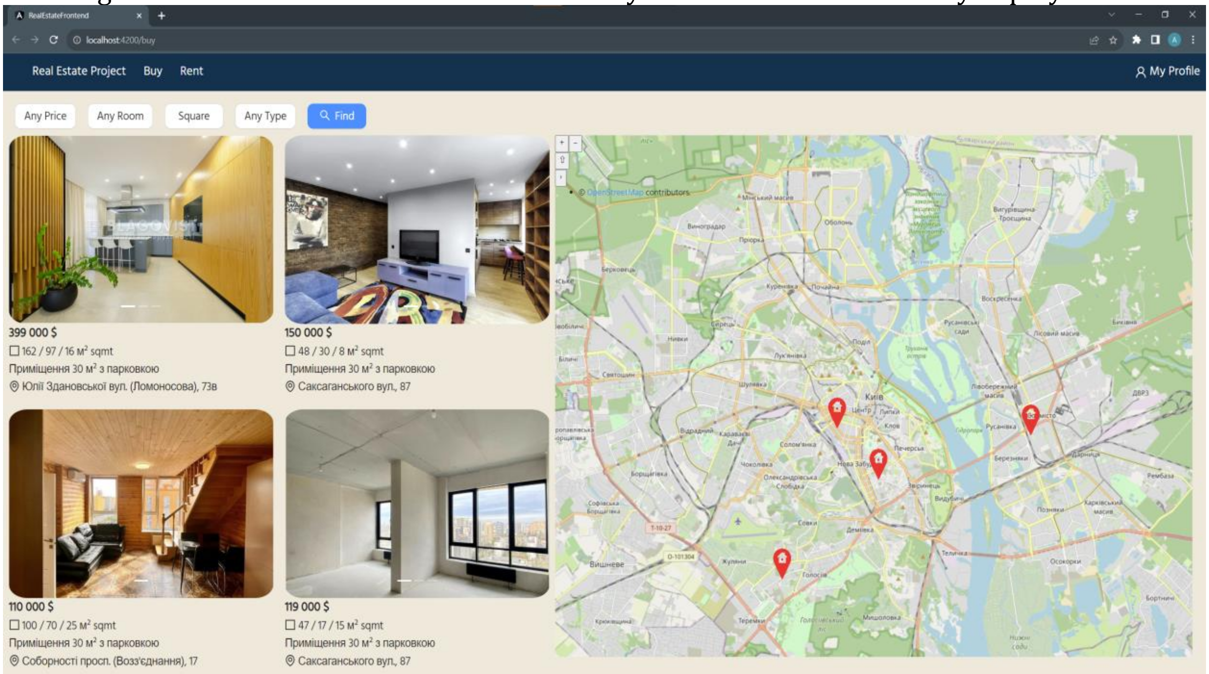
Backend service. Responsible for the business logic of the application - transfers data to the "Data providing service" with the parameters received from the user, in order to determine the value of the real estate. Also provides various functions for data analysis and processing, real estate recommendations, email notification functions, data caching and user authorization. Another task of the service is that it provides the ability to work with all database entities.

Frontend service. And finally, the service responsible for the interaction of the program with the end user. It visualizes data in all possible ways, presenting it in an easy-to-understand form and providing an intuitive application experience. Also, a requirement for the service is the correct reception and transmission of data for communication with the rest of the services.

The implementation of the "Data providing service" uses the Python programming language, as well as the libraries Pandas, Pyspark ML, Scikit-learn, OpenStreetMap Nominatim API, Python Overpass API. The server part of the "Backend service" application was developed using the Java programming language and its Spring Boot framework. HTML, CSS technologies, as well as the JavaScript programming language and one of its most popular frameworks, Angular, were used to develop the user part of the "Frontend service".

The system deployment is done by containerizing each individual service using Docker tools, and connected to each other using Docker Compose.

In Figure 7 and 8 are shown the results of the system that was successfully deployed.



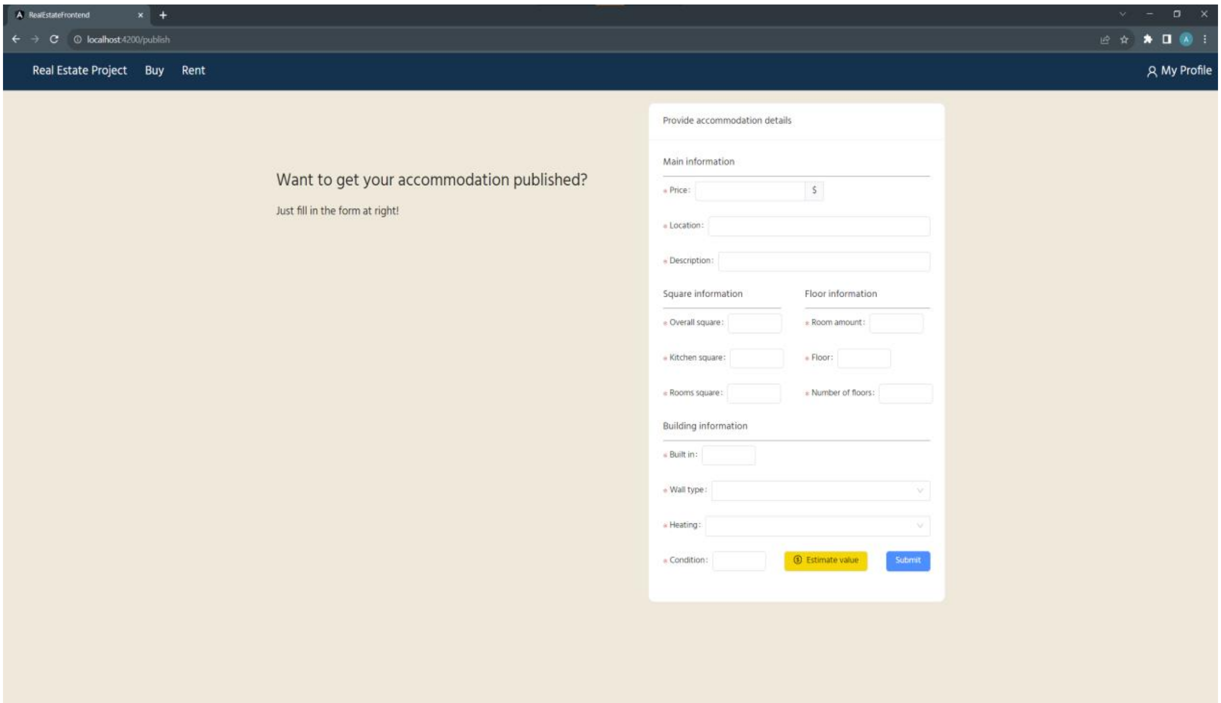
399 000 $
□ 162 / 97 / 16 м² sqmt
Приміщення 30 м² з парковкою
Юлії Здановської вул. (Ломоносова), 73в

150 000 $
□ 48 / 30 / 8 м² sqmt
Приміщення 30 м² з парковкою
Саксаганського вул., 87

110 000 $
□ 100 / 70 / 25 м² sqmt
Приміщення 30 м² з парковкою
Соборності просп. (Возз'єднання), 17

119 000 $
□ 47 / 17 / 15 м² sqmt
Приміщення 30 м² з парковкою
Саксаганського вул., 87

Figure 7: Display of the real estate value estimate for the buyer



Want to get your accommodation published?

Just fill in the form at right!

Figure 8: The result of real estate evaluation according to the parameters set in the system

It was verified that all services work correctly and successfully interact with each other. Data is correctly transmitted, stored and displayed.

6. Conclusions

The result of the research is the creation of a regression model of machine learning, which can be used to determine the value of real estate, depending on its physical parameters and geographical location, as well as a developed information system that provides the opportunity to fully use all the capabilities of this model. The peculiarity of the system is that in the process of training the model, not only the basic characteristics of real estate are used, but also geospatial data. That is, the presence and number of locations of a certain type in a given radius from the real estate object is taken into account, which allows to predict the cost with higher accuracy.

The process of creating a machine learning model is conventionally divided into four stages, which include data collection, filtering, processing, addition, division into different samples and training the model based on this data.

The quality of the machine learning model was checked on the validation sample of data, the forecast results and the absolute error were visualized using graphs and charts. After analyzing this information, it was concluded that the results of the model meet the requirements of the system, and therefore the research was successful.

It was decided to divide the system into three separate services, each of which will be responsible for a certain list of functions. The programming languages, libraries and frameworks that should be used in the development of each service are substantiated. The purpose of the services is described, as well as the system architecture is schematically visualized, the main functions of each service and the connections between them are shown. System deployment is done by containerization and configuration of created containers using Docker capabilities.

We can single out the following positive effects that the implementation of the system will provide:

- *Simplifying the search for real estate*: the system will allow buyers to find real estate that matches their needs and capabilities faster and more efficiently.

- *Improvement of cost estimation*: real estate sellers will be able to more accurately estimate the value of their property, which will help them find buyers and make profitable deals.
- *Reduction of risk for investors*: investors will be able to use the system to analyze and select real estate objects with higher potential.
- *Positive impact on the real estate market*: all interested stakeholders (buyers, sellers, agents and investors) will receive tools that will help improve the efficiency of the real estate market and facilitate interaction between them.

This information system should simplify and speed up the process of finding the optimal real estate for purchase and evaluating its value for further sale. Implementation of the system in the field of real estate will be useful for buyers and sellers of real estate, as well as for agents and investors, simplifying their business processes.

References

[1] The Importance of Accurate Property Valuation in Real Estate, 2023. URL: https://sugermint.com/the-importance-of-accurate-property-valuation-in-real-estate/.

[2] AI in real estate property valuation: Is it really a game-changer?, 2023 URL: https://mdevelopers.com/blog/ai-real-estate-property-valuation.

[3] I. Kolesnikova, Using Artificial Intelligence for Real Estate: A Comprehensive Guide, 2023. URL: https://mindtitan.com/resources/industry-use-cases/artificial-intelligence-in-real-estate/.

[4] REAL-TIME PROPERTY VALUATIONS: HOW AI ALGORITHMS ARE MAKING IT POSSIBLE, 2023. URL: https://www.realspace3d.com/blog/real-time-property-valuations-how-ai-algorithms-are-making-it-possible/.

[5] O. Veres, P. Ilchuk, O. Kots, Data Science Methods in Project Financing Involvement, International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2021, Vol. 2, pp. 411–414. doi: 10.1109/CSIT52700.2021.9648679.

[6] O. Veres, P. Ilchuk and O. Kots, Data Analytics on Debt Financing Research Based on Scopus and WoS Metrics, In 2023 IEEE 18th International Conference on Computer Science and Information Technologies (CSIT), Lviv, Ukraine, 2023, doi: 10.1109/CSIT61576.2023.10324179.

[7] DIM.RIA – all real estate of Ukraine. Sale and rent of any real estate, 2024. URL: https://dom.ria.com/uk/.

[8] OLX.ua ads: Ukrainian classifieds service, 2024. URL: https://www.olx.ua/uk/nedvizhimost/.

[9] Zillow.Com, Agents. Tours. Loans. Homes, 2024. URL: https://www.zillow.com/.

[10] Realtor.Com, Homes for Sale, Real Estate & Property Listing, 2024. URL: https://www.realtor.com/.

[11] Redfin.Com, Real Estate, Homes for Sale, MLS Listings, Agents, 2024. URL: https://www.redfin.com/.

[12] PropStream.Com, Most Trusted Provider of Real Estate Information, 2024. URL: https://www.propstream.com/.

[13] N. Berezhna, Buying a home: do you need a realtor and how much do his services cost in Ukraine, 2021. URL: https://realestate.24tv.ua/kupivlya-zhitla-potriben-rieltor-skilki-koshtuyut-ostanni-novini_n1525065.

[14] Zillow.Com, How much is my home worth?, 2024. URL: https://www.zillow.com/how-much-is-my-home-worth/.

[15] D. Castillo, Machine Learning Regression Explained, 2021. URL: https://www.seldon.io/machine-learning-regression-explained.

[16] Gelman, Andrew, and Jennifer Hill. Data analysis using regression and multilevel/hierarchical models. Cambridge university press, 2006. Gelman, A.; Hill, J. Data

Analysis Using Regression and Multilevel/Hierarchical Models; Cambridge University Press: New York, NY, USA, 2006.

[17] Finch, W. Holmes, Jocelyn E. Bolin, and Ken Kelley. Multilevel modeling using R. Crc Press, 2019.

[18] Evans, Clare R., George Leckie, and Juan Merlo. "Multilevel versus single-level regression for the analysis of multilevel information: the case of quantitative intersectional analysis." Social Science & Medicine 245 (2020) 112499.

[19] What is Simple Linear Regression in Machine Learning?, 2023. URL: https://www.simplilearn.com/what-is-simple-linear-regression-in-machine-learning-article.

[20] Maulud, Dastan, and Adnan M. Abdulazeez. "A review on linear regression comprehensive in machine learning." Journal of Applied Science and Technology Trends 1.4 (2020) 140-147.

[21] D. Polzer, 7 of the Most Used Regression Algorithms and How to Choose the Right One, 2021. URL: https://towardsdatascience.com/7-of-the-most-commonly-used-regression-algorithms-and-how-to-choose-the-right-one-fc3c8890f9e3.

[22] Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. John Wiley & Sons, 2021.

[23] C. Dawson, Understanding Multiple Linear Regression, 2021. URL: https://medium.com/swlh/understanding-multiple-linear-regression-e0a93327e960.

[24] Mahaboob, B., et al. "A study on multiple linear regression using matrix calculus." Advancecs in Mathematics Scientifc journal 9.7 (2020): 1-10.

[25] S. Bouzebda, Y. Souddi, F. Madani, Weak Convergence of the Conditional Set-Indexed Empirical Process for Missing at Random Functional Ergodic Data. Mathematics 12 (2024). https://doi.org/ 10.3390/math12030448

[26] Y. Zhou, D. He, Multi-Target Feature Selection with Adaptive Graph Learning and Target Correlations. Mathematics 12 (2024). https://doi.org/10.3390/math12030372

[27] T. Li, K. A. Frank, M. Chen, A Conceptual Framework for Quantifying the Robustness of a Regression-Based Causal Inference in Observational Study. Mathematics 12 (2024). https://doi.org/10.3390/math12030388

[28] Leyland, Alastair H., and Peter P. Groenewegen. Multilevel modelling for public health and health services research: health in context. Springer Nature, 2020.

[29] Measure of impurity, 2019. URL: https://medium.com/@viswatejaster/measure-of-impurity-62bda86d8760.

[30] Bentéjac, Candice, Anna Csörgő, and Gonzalo Martínez-Muñoz. "A comparative analysis of gradient boosting algorithms." Artificial Intelligence Review 54 (2021) 1937-1967.

[31] P. Rishit, Understanding K-Nearest Neighbors: A Simple Approach to Classification and Regression, 2023. URL: https://pub.towardsai.net/understanding-k-nearest-neighbors-a-simple-approach-to-classification-and-regression-e4b30b37f151.

[32] Chicco, Davide, Matthijs J. Warrens, and Giuseppe Jurman. "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation." PeerJ Computer Science 7 (2021) e623.

[33] P. Bhandari, Central Tendency | Understanding the Mean, Median & Mode, 2023. URL: https://www.scribbr.com/statistics/central-tendency/ .

[34] What Is R Squared And Negative R Squared, 2018. URL: http://www.fairlynerdy.com/what-is-r-squared/ .

[35] ML | Introduction to Data in Machine Learning, 2023. URL: https://www.geeksforgeeks.org/ml-introduction-data-machine-learning/.

[36] What are the average prices on the secondary housing market in Ukraine: how much will you have to pay for a one-room apartment, 2023. URL: https://sud.ua/uk/news/ukraine/259841-kakie-srednie-tseny-na-vtorichnom-rynke-zhilya-po-ukraine-skolko-pridetsya-otdat-za-odnokomnatnuyu-kvartiru.

[37] Real estate Kyiv and region. 2024. URL: https://t.me/ppbestate.

[38] Real estate of the Kyiv region, 2024. URL: https://t.me/Neruhomist_Kyiv_region.

[39] Telethon’s Documentation, 2024. URL: https://docs.telethon.dev/en/stable/

[40] Nominatim 4.3.0 Manual, 2024. URL: https://nominatim.org/release-docs/latest/api/Overview/.

[41] Python Overpass API, 2024. URL: https://python-overpy.readthedocs.io/en/latest/.

[42] L. Silva, Why Spring Boot is so popular: all about the framework, 2023. URL: https://www.linkedin.com/pulse/why-spring-boot-so-popular-all-framework-leonardo-holanda-e-silva.

[43] H. Dhaduk, Angular vs React: Which to Choose for Your Front End in 2023? 2023. URL: https://www.simform.com/blog/angular-vs-react/.

[44] O. Veres, N. Kunanets, V. Pasichnyk, N. Veretennikova, R., Korz, A. Leheza, Development and Operations-the Modern Paradigm of the Work of IT Project Teams. In 2019 IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT) , (2019, September), 3, 103-106. doi: 10.1109/STC-CSIT.2019.8929861.

[45] O. Veres, P. Ilchuk, O. Kots, Y. Levus, O. Vlasenko, Recommendation System for Leisure Time-Management in Quarantine Conditions, CEUR Workshop Proceedingsthis Vol-3312 (2022) 263–282.