System of modeling and forecasting real estate prices based on machine learning methods

Irina Kalinina[†], Aleksandr Gozhyj[*,†], Viktoria Chorna[†], Victor Gozhyi[†] and Sergii Shiyan[†]

Petro Mohyla Black Sea National University, St. 68 Desantnykiv, 10, Mykolaiv, 54000, Ukraine

Abstract

The article examines how the problem of forecasting real estate prices is solved using a systematic approach to modeling and forecasting. Machine learning methods were systematically used to solve the problem. The systematic approach to modeling and forecasting is based on the analysis of the studied processes, establishing the types of existing characteristic uncertainties, assessing the structure and parameters of the model, as well as forecasts based on the constructed model. It combines three groups of tasks on a single methodological basis: the task of data analysis and pre-processing; the task of building models and their evaluation; the task of building forecasts and their evaluation. The structure of the systematic approach to modeling and forecasting is developed and presented. An important aspect that affects the effectiveness of using machine learning methods is the process of pre-processing data. Improving the methods of pre-processing data is a complex task that must be solved systematically, taking into account the specifics of the real estate market. Therefore, in this study, considerable attention is paid to the process of pre-processing data and research aimed at increasing the effectiveness of predictive values. The architecture of an information system for solving modeling and forecasting problems is developed and presented. As an example of implementing an information system, the problem of forecasting real estate prices is considered. The results of the following stages are presented: data collection, research and data preparation, model training on data, determining model efficiency, improving the efficiency of basic models. The following groups of models were used to solve the forecasting problem: regression models, tree models. The effectiveness of forecasting solutions was assessed using the MAE, MSE, RMSE, MAPE metrics. To improve the quality of forecasts, a single-layer structure of a heterogeneous ensemble of models based on stacking is proposed.

Keywords

systems approach to modeling and forecasting, real estate price forecasting, information system, uncertainties, machine learning

1. Introduction

Solutions to many applied machine learning problems depend on various factors: the specifics of the subject area and the structure of the initial data set, the volume of data, the presence of various types of uncertainties in the data. But one of the main tasks solved in machine learning problems is to obtain accurate predictions about the behavior of complex objects and systems. Predictive values are obtained based on preliminary data analysis and analysis of the past behavior of the system under study. Often, when determining predictive values, many problems arise that cannot be solved by known methods and appropriate algorithms. Problems arise because sometimes the mechanisms of real data generation are not precisely known or the sample size is insufficient to build a high-quality predictive model. Real data often contain nonlinearities and/or non-stationarity of various types. This requires careful analysis and pre-processing of data because the quality of pre-prepared data significantly affects the quality of the predictive model. A predictive model built using machine learning methods significantly depends on the data pre-processing process because data uncertainties are identified and taken into account at the stages of this process.

[*] *CIAW-2025: Computational Intelligence Application Workshop, September 26-27, 2025, Lviv, Ukraine*

[1*] Corresponding author.

[†] These authors contributed equally.

irina.kalinina1612@gmail.com (I. Kalinina); alex.gozhyj@gmail.com (A. Gozhyj); chornav2008@gmail.com (V. Chorna); gozhyi.v@gmail.com (V. Gozhyi); shyiansi@gmail.com (S. Shiyan)

0000-0001-8359-2045 (I. Kalinina); 0000-0002-3517-580X (A. Gozhyj); 0000–0002–6205–7163 (V. Chorna); 0000–0002–5341–0973 (V. Gozhyi); 0000-0001-9255-9511 (S. Shiyan)

The use of modern methodology of systems analysis in solving modeling and forecasting problems is necessary for building more accurate forecasting models. This allows using mathematical models for modeling processes of various nature based on modern developments in the field of probabilistic statistical methods and estimation theory [1-3].

Most modern forecasting methods [4-7] are not used systematically, therefore, it does not allow to obtain better estimates of forecasts in the presence of uncertainties of different types [3,8-10]. Independent use of different forecasting methods significantly reduces the efficiency of solving modeling and forecasting problems. When using methods of analysis and pre-processing of data to solve machine learning problems, there are limitations associated with the presence of various types of uncertainties in the input data. They depend on a number of factors and do not allow to make appropriate assumptions and establish laws of distribution of uncertain features, and to draw conclusions about the influence of individual input values on the result. In the tasks of preliminary data analysis, there are various types of uncertainties, such as imprecision and uncertainty of various parameters in the data, insufficient information about the data distribution, nonlinearity, non-stationarity, and stochasticity of the processes under study.

Analysis of real data usually requires taking into account various types of data uncertainties, as well as the structure of the process under study, uncertainties in model parameters, and uncertainties related to the quality of models and forecasts. All types of uncertainties can be divided into *statistical*, *structural*, and *parametric* [1,3,9,10].

Statistical uncertainty is caused by the data itself, i.e., the presence of omissions, anomalous values of features, the presence of data repetitions, measurement errors, a small sample size of data, and the influence of external random disturbances on the process under study. Taking into account various types of statistical uncertainty during analysis and pre-processing of data when modeling and forecasting real data allows increasing the accuracy and efficiency of predictive models [11,12].

Structural uncertainty arises when evaluating the structure of the model based on data because the structure of the studied process is unknown or not clearly defined. For example, when using a functional approach to building a model, the structure of the object (or process) is usually unknown. The model structure is estimated using appropriate methods: correlation analysis, lag estimation, testing for nonlinearity and non-stationarity, mutual information estimation, detection of external disturbances, etc. At each stage, the corresponding estimates are obtained, which are random variables. This adds uncertainty to the final result [13].

Parametric uncertainty is a consequence of the presence of statistical and structural uncertainties. The approximation of model structure estimates, the presence of external disturbances, measurement errors, and the inability to establish the correct type of data distribution lead to a bias (shift) of model parameter estimates from the exact values and an increase in the dispersion of these estimates. Therefore, it is important to apply a systematic approach to the selection of methods for estimating model parameters that are built on real data [14,15].

One of the machine learning tasks, which is characterized by variability, data complexity and various types of uncertainties, is the task of forecasting real estate prices. The main feature of real estate is that it is the largest asset class whose value increases over time. Real estate is both a consumer and an investment product. The most important property of real estate is that it constitutes a significant part of all assets for the majority of the population. At the same time, an important task is real estate valuation – this is the process of developing a fair and acceptable market value of real estate for both the buyer and the seller. This process is a complex systemic task that depends on many environmental, physical and macroeconomic factors and variables. Another feature is that the real estate sector is a rapidly changing, competitive and opaque sector, where access to real information is difficult. Therefore, in such conditions, data mining methods can be a source of information for many stakeholders and be used as an effective tool for responding to changing conditions. Therefore, the development of systems that make accurate price forecasts according to the real estate being purchased is relevant and of great importance [16].

In addition, in order to accurately predict price changes, both individuals and companies need to know the current and actual value of any property [17]. Therefore, there is a growing need to develop real estate valuation models to obtain accurate real estate price forecasts in order to avoid subjectivity and bias in real estate valuation [18,19]. In this context, works [20-22] provide a comprehensive analysis of regression types for machine learning models and deep learning models, which have not been widely used in real estate valuations, but provide effective results in predicting real estate prices.

In works [23,24], examples of analysis and evaluation of different real estate lenses are given. These analysis examples are complex and cover a variety of issues that require multidimensional and more accurate determination of market value.

Today, research on machine learning and deep learning is accelerating developments in this field and spreading the use of machine learning methods in various fields. In this context, there is research on determining real estate prices. In particular, the presence of too many parameters in determining real estate prices makes machine learning and deep learning models particularly attractive in this field. In [25], a system for accurate forecasting of real estate prices based on machine learning algorithms: linear regression, random forest, boosted regression and artificial neural networks was presented.

In [21,26], the results of various machine learning methods were presented, which identified the advantages and disadvantages of each method. The results of the study showed that the most effective models are always ensemble models, based on trees and regression.

In [27,28] it is shown that machine learning methods such as XGBoost, which are not often used in this field, can be a better alternative to methods such as artificial neural networks and traditional multiple regression analysis, which are often preferred, especially in real estate price forecasting problems. The XGBoost model has demonstrated efficiency compared to other models used in the study. Although there is no significant difference between the results obtained by the XGBoost model and the neural network model in the study, there is a significant difference between linear, lasso and comb regression.

In the study [29] it is shown that the efficiency of solving the problem depends on the sample size. For example, neural networks give better results with large sample sizes. In [30] it is also shown that the efficiency of fuzzy neural networks in predicting real estate prices directly depends on the quality of the data used.

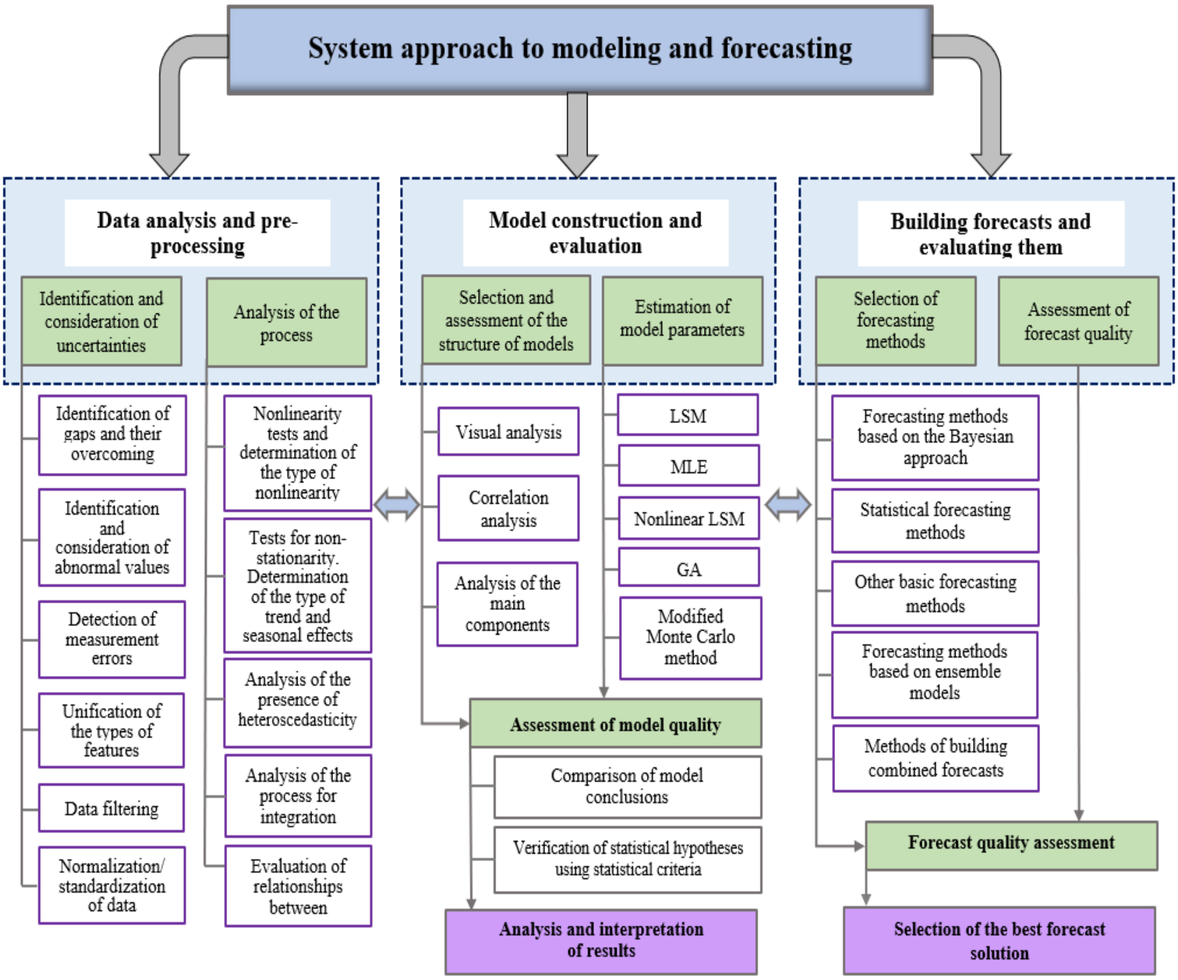
The aim of this study is to develop methods for modeling and forecasting real estate prices using machine learning methods. One of the most important aspects that affect the success of using machine learning methods is the process of data pre-processing. Improving data pre-processing methods is a complex task that must be solved on the basis of a systematic approach taking into account the specifics of the real estate market. Therefore, in this study, special attention is paid to the process of data pre-processing and research aimed at improving the effectiveness of predictive values.

Problem statement. The purpose of this article is to build an information system for forecasting real estate prices based on the systematic use of machine learning methods. To do this, it is necessary to determine the main features of a systematic approach to modelling and forecasting processes. To build and implement a generalised algorithm for data analysis and pre-processing. To develop and investigate the architecture of an information system for solving the problem of forecasting real estate prices, as well as to experimentally present the advantages of a systematic approach for solving machine learning problems.

2. Models and methods

2.1. System approach to modeling and forecasting

The system approach is a methodological basis for solving modeling and forecasting problems. The basis of the system approach is the consistent and interconnected use of groups of methods for analyzing and pre-processing data, methods for modeling and assessing the quality of models, and methods for forecasting and assessing the quality of the obtained forecast values. This process is iterative and hierarchical. The structural diagram of the system approach for solving modeling and forecasting problems in machine learning problems is presented in Fig. 1.



System approach to modeling and forecasting
Data analysis and pre-processing
Identification and consideration of uncertainties
Analysis of the process
Identification of gaps and their overcoming
Identification and consideration of abnormal values
Detection of measurement errors
Unification of the types of features
Data filtering
Normalization/ standardization of data
Nonlinearity tests and determination of the type of nonlinearity
Tests for non-stationarity. Determination of the type of trend and seasonal effects
Analysis of the presence of heteroscedasticity
Analysis of the process for integration
Evaluation of relationships between
Model construction and evaluation
Selection and assessment of the structure of models
Estimation of model parameters
Visual analysis
Correlation analysis
Analysis of the main components
LSM
MLE
Nonlinear LSM
GA
Modified Monte Carlo method
Assessment of model quality
Comparison of model conclusions
Verification of statistical hypotheses using statistical criteria
Analysis and interpretation of results
Building forecasts and evaluating them
Selection of forecasting methods
Assessment of forecast quality
Forecasting methods based on the Bayesian approach
Statistical forecasting methods
Other basic forecasting methods
Forecasting methods based on ensemble models
Methods of building combined forecasts
Forecast quality assessment
Selection of the best forecast solution

Figure 1: Structure of a systems approach to modeling and forecasting.

The system approach is based on the analysis of a complex process (object) that is being studied. This system methodology begins with the identification and consideration of uncertainties, primarily of the statistical type. The cleaned data after analysis and pre-processing are used to build basic models in which uncertainties of the structural and parametric types are identified and taken into account. The forecast values obtained at the forecasting stage can be improved, if necessary, by combining forecast values or by using heterogeneous ensembles of models [1,3,9].

The systematic methodology of modeling and forecasting solves the following tasks:

- Using methods of analysis and pre-processing of data in accordance with the machine learning task and the characteristics of the data set.

- Data analysis, identification and consideration of statistical uncertainties (gaps, anomalous values, errors, repetitions in the data, study of the type of distribution of features).
- Identification and overcoming of structural and parametric uncertainties in the modeling process.
- Comprehensive assessment of the adequacy of models and the quality of forecasts using a set of criteria.
- Construction and analysis of basic alternative forecasting models.
- Systematic approach to the selection of methods for estimating the parameters of forecasting models (LSM, MLE, Nonlinear LSM and others).
- Optimization of the structures of basic forecasting models.

Thus, the systematic approach to modeling and forecasting combines three groups of tasks and methods: tasks of analysis and pre-processing of data; tasks related to the construction of basic forecasting models and their evaluation; tasks of constructing forecasts and assessing their quality.

Each of these groups of tasks combines methods and approaches that constitute elements of information technology. The first task of data analysis and pre-processing is divided into two subgroups of methods: methods of identifying and taking into account various types of statistical uncertainties and methods of analysing the process under study and its individual components. The second task of building basic forecast models and assessing their adequacy is divided into two subgroups of methods: methods of selecting and evaluating the structure of models and methods of estimating model parameters. The third task of building forecasts and assessing their quality is also divided into two subgroups of methods: methods and approaches to building forecast values and methods of evaluating them. All methods and approaches to solving machine learning problems are used systematically and inter-connectedly. Information technologies for solving real data analysis problems and solving various machine learning problems are built on the basis of a systematic approach.

2.2. Information system for modeling and forecasting

Based on the structure of the system approach to solving modelling and forecasting problems, which is presented in Fig. 1, the general structure of the forecasting information system based on the use of machine learning methods has been developed. The system consists of a sequential implementation of subsystems: an information storage subsystem, a data analysis and pre-processing subsystem, a modelling subsystem, and a forecasting subsystem. The structure of the modelling and forecasting information system is presented in Fig. 2. The system combines groups of methods into subsystems according to the main tasks of the system approach. In the generalised presented data analysis and pre-processing subsystem, the procedures for identifying and processing missing values in the data, identifying and processing anomalous values, as well as the procedures for filtering, smoothing, feature selection, and their normalization are implemented.

The modelling subsystem of the information system presents a data set distribution block and two samples (training and test), procedures for building basic forecasting models and procedures for assessing the adequacy of models. The forecasting subsystem presents a procedure for building forecast values based on basic forecasting models and a procedure for assessing the quality of forecasts. This subsystem provides a procedure for improving forecast values using an ensemble approach. The ensemble approach involves building single-layer or multi-layer heterogeneous ensembles of forecasting models using bagging, boosting, and stacking methods. An information system for solving modelling and forecasting problems based on real data is the result of the systematic use of modelling and forecasting methods and approaches.

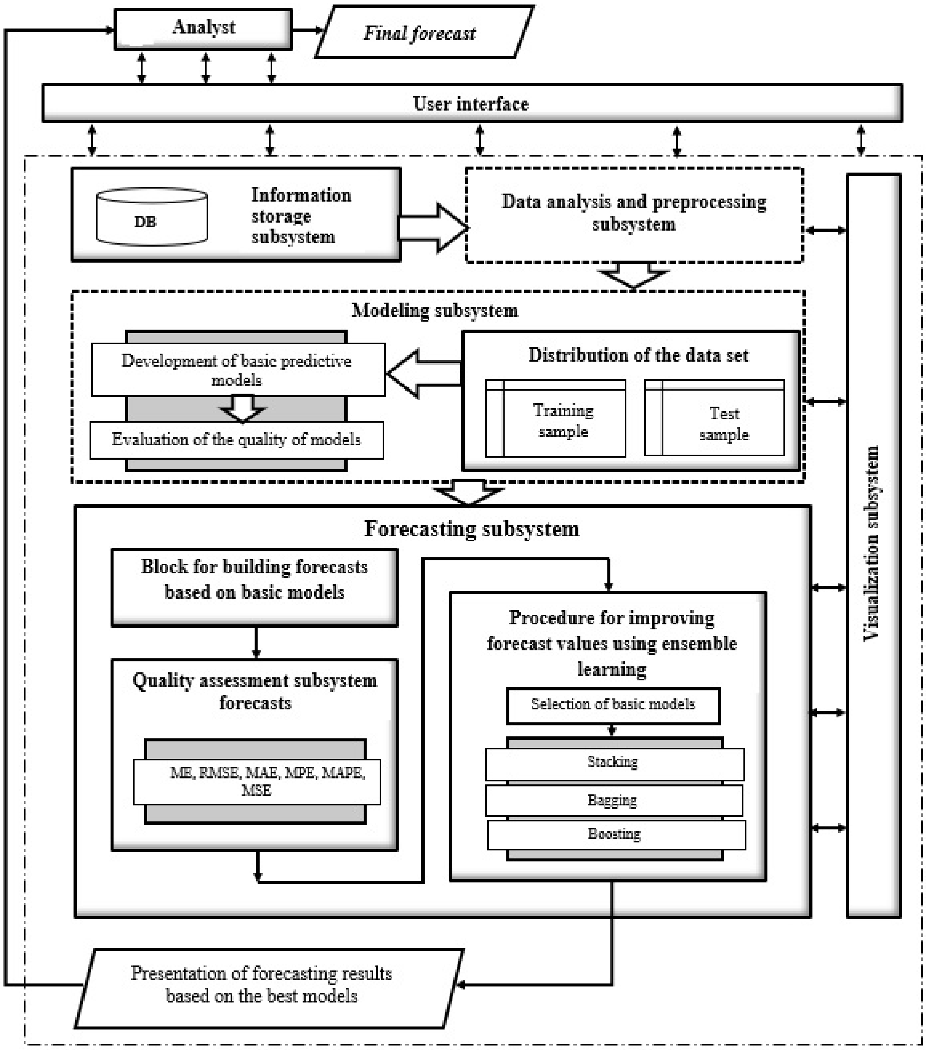

Analyst
Final forecast
User interface
DB
Information storage subsystem
Data analysis and preprocessing subsystem
Modeling subsystem
Development of basic predictive models
Evaluation of the quality of models
Distribution of the data set
Training sample
Test sample
Forecasting subsystem
Block for building forecasts based on basic models
Quality assessment subsystem forecasts
ME, RMSE, MAE, MPE, MAPE, MSE
Procedure for improving forecast values using ensemble learning
Selection of basic models
Stacking
Bagging
Boosting
Visualization subsystem
Presentation of forecasting results based on the best models

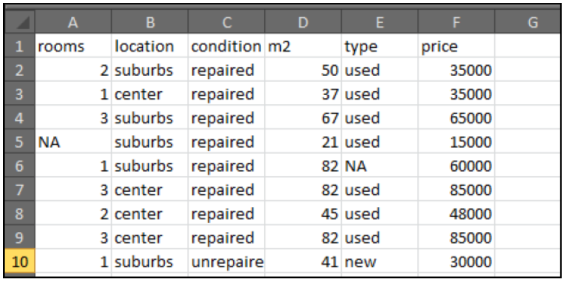
Figure 2: Structure of the modeling and forecasting information system.

3. Experimental part

3.1. Data analysis and pre-processing

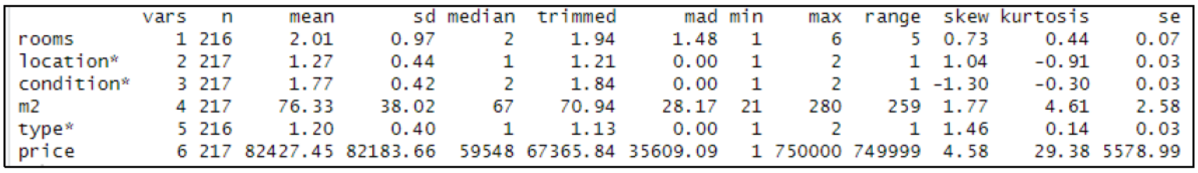
To demonstrate the advantages of a systematic approach to solving modeling and forecasting problems, the *flats.csv* [31] dataset was used, which contains information about real estate in the form of a certain set of characteristics. The file lists apartment prices, type, square footage, condition, location, and number of rooms (Fig. 3.).

During the statistical description, the main indicators were calculated for each variable, allowing to analyse their distribution and variability. These indicators include the mean, median, standard deviation and other parameters that help to identify data features and possible anomalies (Fig. 4).



	A	B	C	D	E	F	G
1	rooms	location	condition	m2	type	price	
2	2	suburbs	repaired	50	used	35000	
3	1	center	repaired	37	used	35000	
4	3	suburbs	repaired	67	used	65000	
5	NA	suburbs	repaired	21	used	15000	
6	1	suburbs	repaired	82	NA	60000	
7	3	center	repaired	82	used	85000	
8	2	center	repaired	45	used	48000	
9	3	center	repaired	82	used	85000	
10	1	suburbs	unrepaire	41	new	30000	

Figure 3: Example data set *flats.csv*.



```
           vars   n     mean       sd median  trimmed      mad min    max  range  skew kurtosis      se
rooms         1 216     2.01     0.97      2     1.94     1.48   1      6      5  0.73     0.44    0.07
location*     2 217     1.27     0.44      1     1.21     0.00   1      2      1  1.04    -0.91    0.03
condition*    3 217     1.77     0.42      2     1.84     0.00   1      2      1 -1.30    -0.30    0.03
m2            4 217    76.33    38.02     67    70.94    28.17  21    280    259  1.77     4.61    2.58
type*         5 216     1.20     0.40      1     1.13     0.00   1      2      1  1.46     0.14    0.03
price         6 217 82427.45 82183.66  59548 67365.84 35609.09   1 750000 749999  4.58    29.38 5578.99
```

Figure 4: Descriptive statistics on variables.

Each of these variables reflects key characteristics of real estate objects that can significantly affect their market value. In particular, it is important to consider that the analysis allows you to identify patterns, as well as assess the degree of influence of various parameters on price formation [1]. Based on the statistical description, it can be concluded that additional processing is necessary before using the data in the forecasting model (Table 1).

Table 1

Descriptive statistics analysis of a data set

No	Variable name	Feature analysis from a dataset
1	rooms	Number of rooms in an apartment. The sample includes 216 apartments. The average number of rooms is 2.01, the minimum is 1, and the maximum is 6.
2	location	Apartment location, categorical variable (value 1 or 2). Total 217 observations. Specifies the geographic location of the home, which may affect its price.
3	condition	Apartment condition, categorical variable. In the sample of 217 observations, the average value is 1.77. Affects the attractiveness of the object for buyers.
4	m2	Apartment area in square meters. 217 apartments are presented. The average area is 76.33 m², the minimum is 21 m², the maximum is 280 m². Area is an important factor that directly affects the price.
5	type	Apartment type, a categorical variable (e. g., new construction or secondary market). The sample contains 216 observations. Apartment type also affects market value.
6	price	Apartment price. There are 217 observations in the sample. The average price is 82,427.45 UAH, with a large range from 1 UAH to 1,750,000 UAH. This is the main indicator for analysis and forecasting.

The analysis revealed the presence of missing values, as well as values that do not meet logical or statistical expectations. These anomalies can negatively affect the accuracy and reliability of the model, since it may perceive them as valid data, which can lead to distortion of the results. Therefore, it is important to implement data cleaning stages, including filling in gaps, correcting inadequate values, or deleting them, to ensure high quality and correctness of the data that will be used for further analysis and modeling.

For the implementation of an information system for modeling and forecasting real estate prices, an important stage is data pre-processing [1,9,10]. This stage provides data preparation for effective training of basic forecasting models, which significantly affects the quality of forecasts. The procedure for data analysis and cleaning is presented in the form of an algorithm flowchart in Figure 5.

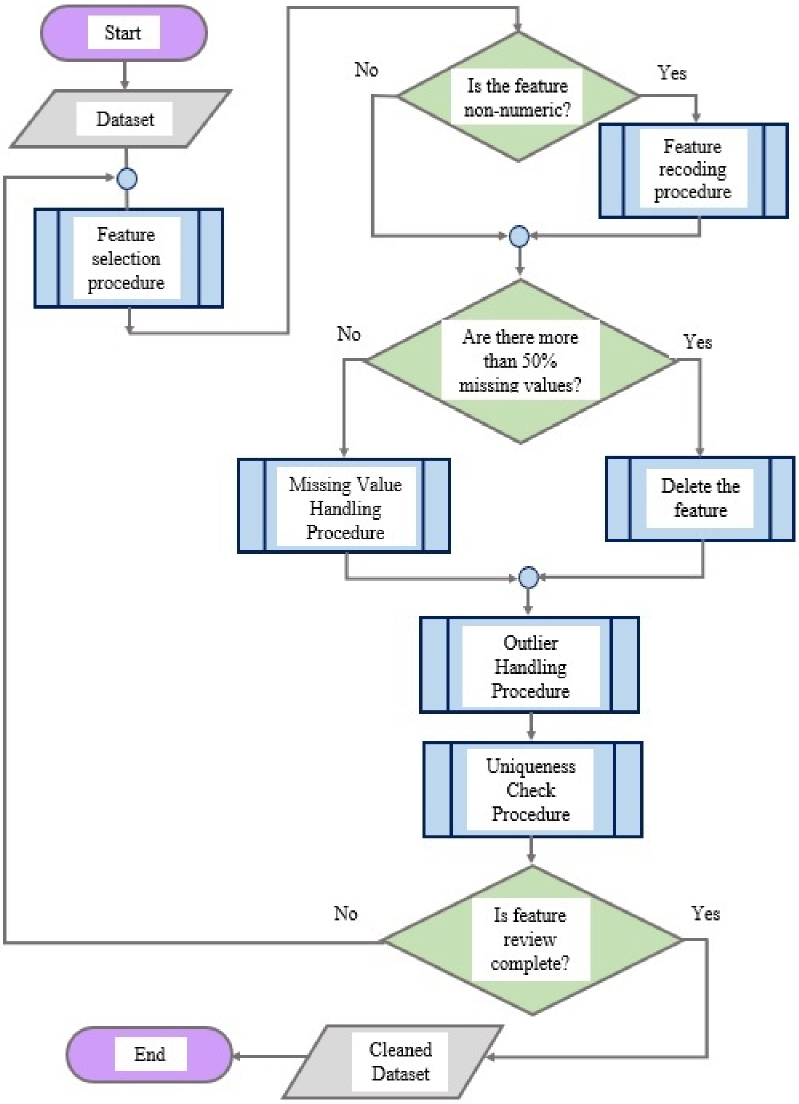

Start
Dataset
Feature selection procedure
No
Is the feature non-numeric?
Yes
Feature recoding procedure
No
Are there more than 50% missing values?
Yes
Missing Value Handling Procedure
Delete the feature
Outlier Handling Procedure
Uniqueness Check Procedure
No
Is feature review complete?
Yes
End
Cleaned Dataset

Figure 5: Block diagram of the data cleaning algorithm.

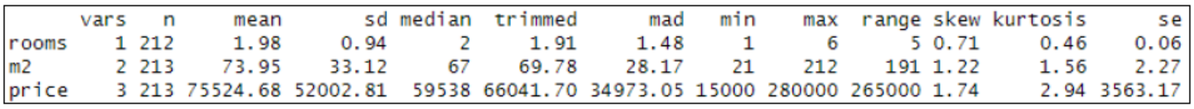
The data cleaning algorithm begins with the stage of loading the dataset. The next step is the selection of features, in which the variables that will be used in further modeling are determined. After this, the type of selected features is evaluated. If the feature is numeric, then the transition to the next stage is performed, where the presence of missing values is checked. If the number of missing values is more than 50%, the feature is removed from the dataset. In the case when the number of missing values is less than 50%, the missing values processing procedure is performed.

Next, the algorithm includes an outlier processing procedure, which involves the detection of anomalous values in the data that may affect the accuracy of the model. After this, a uniqueness check procedure is performed, which includes the detection of duplicate features in the dataset.

The last stage is the evaluation of the completion of feature verification. If all selected features have been verified, the algorithm completes its work, and the cleaned dataset is stored for use in the modeling process. Thus, this algorithm ensures the reliability and accuracy of data, which are important for creating an effective information system.

In the data processing process, it is important to perform filtering in order to focus on observations that are relevant for further analysis. First, all apartments with a price exceeding 300,000 were removed from the table. This allows us to eliminate excessively expensive objects that can distort the results of the analysis. Then, additional filtering was performed, which left only those observations where the apartment price exceeds 10,000. This also helped to remove apartments with an abnormally low cost that do not correspond to market realities.

After filtering, a statistical description of the filtered data was performed for key variables: number of rooms, apartment area in square meters and price. This allowed us to obtain summary statistics for these three variables after cleaning the data set. Thanks to the steps taken, the observation table was significantly cleaned. The number of objects in the sample decreased from 217 to 213, since all observations that did not meet the filtering criteria were removed (see Fig. 6). This improves data quality and ensures the correctness of further analysis and modeling.



```
      vars   n     mean       sd median  trimmed      mad   min    max  range skew kurtosis      se
rooms    1 212     1.98     0.94      2     1.91     1.48     1      6      5 0.71     0.46    0.06
m2       2 213    73.95    33.12     67    69.78    28.17    21    212    191 1.22     1.56    2.27
price    3 213 75524.68 52002.81  59538 66041.70 34973.05 15000 280000 265000 1.74     2.94 3563.17
```

Figure 6: Descriptive statistics on variables after the filtering process.

To visually analyse the variables in the data set, histograms were created that allow us to visually assess the distribution of values for each variable. This allows us to more quickly understand the structure of the data and identify key trends, such as the frequency of occurrence of different values, dominant categories, and potential anomalies. Histograms serve as an effective visualisation tool, making it easier to interpret the results and prepare for further analysis.

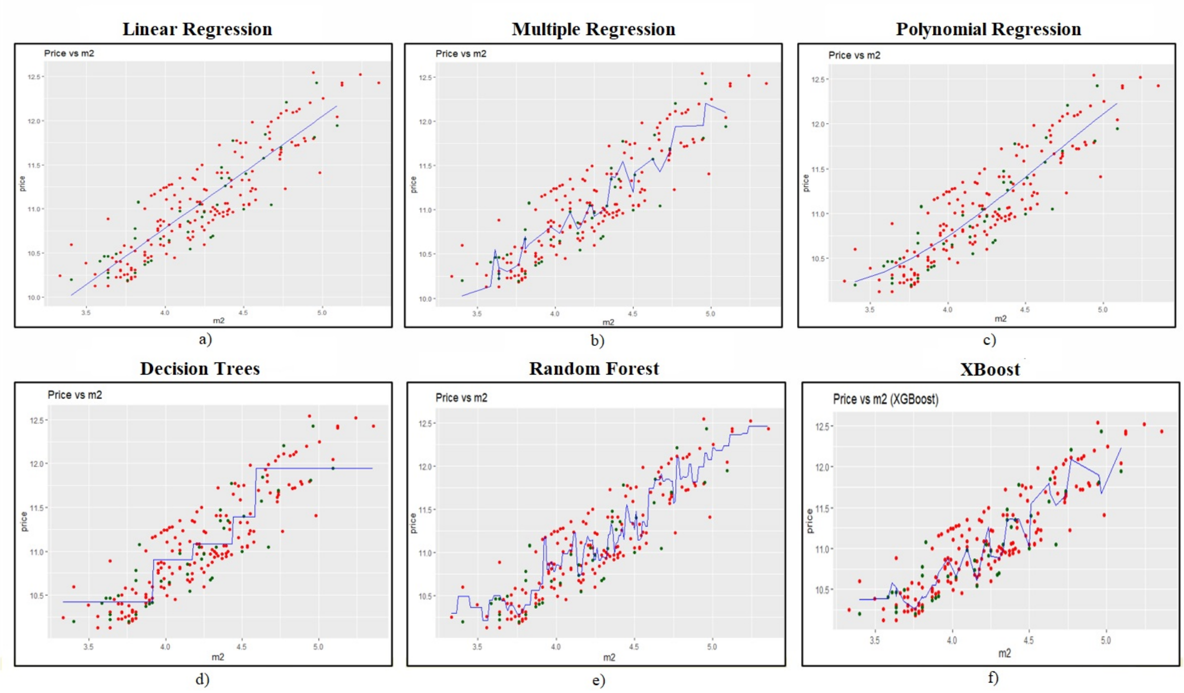
To ensure the correctness and reliability of the data analysis, the data set was checked for missing values in columns containing important information, in particular in the variable's rooms (number of rooms) and type (type of apartment). After that, it was decided to remove rows from the data set containing missing values, in particular in the column's *rooms* (number of rooms) and *type* (type of apartment). Missing values in the resulting attribute *price* were processed by replacing them with average values, which allows you to avoid problems with insufficient data and preserve valuable information. This is an important step in data preparation, since missing values can significantly affect the results of the analysis and, subsequently, the accuracy of the forecasting model. Categorical variables were converted to numerical values, which ensured compatibility with machine learning methods. In the next stage of data analysis, the variables were logarithmically transformed, which allowed stabilising the data variance.

To analyse the influence of individual attributes on the resulting attribute, a correlation analysis was performed. The obtained results of the correlation analysis demonstrate significant relationships between the studied variables, which can be the basis for further analysis and modelling of real estate prices.

3.2. Modeling and forecasting

According to the accepted practice, the cleaned dataset prepared for modeling was split in the ratio of 80:20 [9,10]. Thus, 80% of the data was separated as a training sample, which is prepared for training the models and analyzing their adequacy. Then 20% of the data was left as a test sample, which is intended to check the quality of the underlying predictive models. This avoids over-training, when the model shows high results on the training data, but has poor performance on new, unknown data.

The basic predictive models considered were the linear regression model, the multiple regression model, the polynomial regression model, the decision tree model, the random forest regression model, and the XGBoos model. For each model, a structure was selected and parameters were found under which the models had the best quality indicators of predictions on the test data set. Thus, the problem of taking into account structural and parametric uncertainties was solved. A graphical representation of the results of modeling and forecasting using the basic predictive models is shown in Fig. 7. The figures for each model demonstrate the dependence of the resulting variable on one feature m2. Table 2 presents the values of the quality metrics after training and testing each of the basic predictive models.



Linear Regression

Multiple Regression

Polynomial Regression

a)

b)

c)

Decision Trees

Random Forest

XBoost

d)

e)

f)

Figure 7: Graphical representation of modeling results using basic predictive models (a – linear regression models, b – multiple regression models, c – polynomial regression models, d - decision tree models, e – random forest models, f – XGBoost models).

Table 2

Generalized table for assessing the adequacy of forecast models and the quality of forecasts

Type of model	Model quality metrics				Forecast quality metrics		
	R^2	DW	AIC	F	MSE	MAPE	Theil
Linear Regression	0,774	0.964	15.632	141.064	0.074	3.379	0.0193
Multiple Regression	0.861	1.021	-5.378	255.777	0.045	1.554	0.0096
Polynomial Regression	0.788	0.933	12.965	152.712	0.071	1.555	0.0096
Decision Trees	0.710	1.153	26.386	100.777	0.098	2.110	0.0138
Random Forest	0.761	1.070	18.030	131.188	0.082	2.115	0.0131
XBoost	0.790	1.107	12.576	154.473	0.068	1.675	0,0111

After analyzing the results from Table 2, among the basic forecasting models, the regression model based on multiple regression and the XGBoost model should be distinguished. With the help of these models, the highest quality forecasting values were obtained.

3.3. Approach to improving predictive values

To reduce the error of forecast values, it is necessary to use approaches that make it possible to simultaneously influence the reduction of variance and bias. Reducing the value of each of these components helps to reduce the overall error, and if it is possible to reduce both bias and variance, then the overall forecast error is maximally reduced and the quality of forecasts is improved. An approach that allows implementing a similar technique is ensemble learning. The structures of model ensembles are divided into two groups [32-34]: structures of homogeneous ensembles and structures of heterogeneous ensemble models. The structure of a homogeneous ensemble uses basic forecast models of the same type, while the structure of a heterogeneous ensemble uses basic forecast models of different types. The main idea of ensemble learning is that ensembles work better than their components when the basic forecast models are not identical (use different principles of model construction). A necessary condition for the usefulness of the ensemble approach is that the basic forecast models must have a significant level of differences that make errors independently of each other [35-38]. The limitations of homogeneous ensemble structures can be overcome by using heterogeneous ensembles. Ensemble construction is usually a two-step process: a set of different base models are generated by running different training algorithms on the training data, and then the generated models are combined into an ensemble. Research shows that the strength of a heterogeneous ensemble is related to the performance of the base predictive models and the lack of correlation between them [39-41]. Therefore, to improve the obtained forecast values, a scheme of the ensemble approach was developed, which is presented in Figure 8.

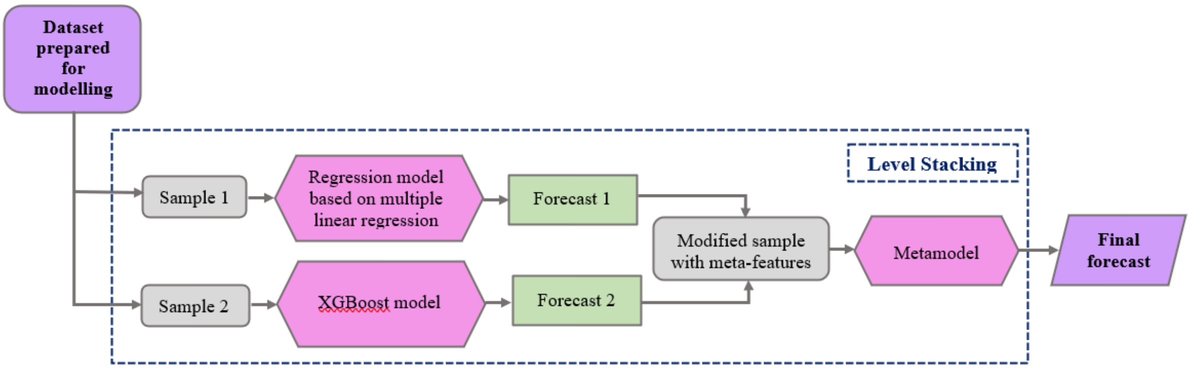

Dataset prepared for modelling
Level Stacking
Sample 1
Regression model based on multiple linear regression
Forecast 1
Modified sample with meta-features
Metamodel
Final forecast
Sample 2
XGBoost model
Forecast 2

Figure 8: Schematic of an ensemble approach to improve forecast values.

Table 3
Table of quality estimates of predictions after the ensemble learning procedure

Type of model	R^2	MSE	MAPE	Theil
Multiple Regression	0,861	0,045	1,554	0,0096
XBoost model	0,790	0,068	1,675	1,0111
Resulting stacking layer	*0,882*	*0,041*	*1,516*	*0,0091*

It is proposed to build a single-layer heterogeneous ensemble of models based on the stacking method with a meta-model based on the support vector method. To improve the quality of aggregation of forecast values, the best models from the modeling stage of the basic forecast models were selected as basic models for stacking: a regression model based on multiple regression and the XBoost model. It is important that the selected models do not correlate with each other. XBoost is also an ensemble model, but XBoost is a homogeneous ensemble structure. Thus, in the scheme presented in Fig. 8, there are two ensemble structures: homogeneous (Boosting) and heterogeneous (Stacking). The estimates of the quality of forecasts after the ensemble training procedure, which are given in Table 3, indicate a decrease in the total error. The stacking model demonstrates the best values of the indicators. This indicates that improving the quality of forecasts through the systematic use of ensemble models gives an advantage compared to the results of forecasts on any basic forecast models.

4. Conclusions

The article presents a systematic approach to modeling and forecasting using the example of the problem of forecasting prices in the real estate market. Machine learning methods were used to solve the problem. The systematic approach is based on the analysis of the studied processes, establishing the types of existing characteristic uncertainties (statistical, structural, parametric), assessing the structure and parameters of the model, as well as forecasts based on the constructed model. A structural diagram of a systematic approach to modeling and forecasting is developed and presented. It combines three groups of tasks on a single methodological basis: data analysis and pre-processing tasks; model building and evaluation tasks; forecast building and evaluation tasks. Particular attention is paid to the data pre-processing process. Improving data pre-processing methods is a complex task that must be solved systematically, taking into account the specifics of the real estate market. Therefore, the article pays significant attention to the data pre-processing process and research aimed at increasing the efficiency of predictive values. The architecture of an information and analytical system for solving forecasting tasks is developed and presented. As an example of solving an applied problem, the process of forecasting prices in the real estate market is considered. The results of the following stages are presented: data collection, research and preparation of data, training the model on data, determining the effectiveness of the model, improving the effectiveness of the model.

The following models were used at the modeling stage: three types of regression models and three types of models built on trees. The effectiveness of forecast solutions was assessed using the quality metrics MAPE, MSE, RMSE and Theil coefficient. To reduce the overall error of the forecast values of the base models, a scheme of a single-layer heterogeneous ensemble of models based on the Stacking method was implemented. The ensemble is built on the best base models of different groups to prevent correlation between them. This approach effectively improves the quality of forecast solutions.

Declaration on Generative AI

The authors did not use any generative AI tools.

References

[1] I. Kalinina, A. Gozhyj, V. Vysotska, E. Malakhov, V. Gozhyj, I. Tregubova, System methodology of data analysis and preprocessing for solving classification problems, in: Proceedings of the 2024 IEEE 19th International Conference on Computer Science and Information Technologies (CSIT), Lviv, Ukraine, 2024. doi:10.1109/CSIT65290.2024.10982630. URL: https://ieeexplore.ieee.org/document/10982630.

[2] V. Andrunyk, A. Vasevych, L. Chyrun, N. Chernovol, N. Antonyuk, A. Gozhyj, V. Gozhyj, I. Kalinina, M. Korobchynskyi, Development of information system for aggregation and ranking of news taking into account the user needs, in: CEUR Workshop Proceedings, vol. 2604, 2020. URL: http://ceur-ws.org/Vol-2604/paper74.pdf.

[3] P. Bidyuk, A. Gozhyj, Z. Szymanski, V. Beglytsia, The methods Bayesian analysis of the threshold stochastic volatility model, in: Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, Ukraine, October 2018, pp. 70–74. doi:10.1109/DSMP.2018.8478474.

[4] S. Marsland, Machine Learning: An Algorithmic Perspective, Massey University, Palmerston North, 2015, 452 p.

[5] V. Lakshmanan, S. Robinson, M. Munn, Machine Learning Design Patterns: Solutions to Common Challenges in Data Preparation, Model Building, and MLOps, 1st ed., O'Reilly Media, 2020, 448 p.

[6] M. H. Veiga, F. G. Ged, Mathematical Foundations of Machine Learning, University of Michigan, 2021, 175 p.

[7] H. Wickham, G. Grolemund, R for Data Science: Import, Tidy, Transform, Visualize, and Model Data, O'Reilly Media, 2017, 520 p.

[8] J. D. Kelleher, B. Mac Namee, A. D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies, 2nd ed., MIT Press, Cambridge, MA, 2020, 798 p.

[9] B. Lantz, Machine Learning with R: Expert Techniques for Predictive Modeling, 3rd ed., Packt Publishing, 2019, 458 p.

[10] A. Nielsen, Practical Time Series Analysis: Prediction with Statistics and Machine Learning, O'Reilly Media, 2019, 504 p.

[11] O. Garasym, L. Chyrun, N. Chernovol, A. Gozhyj, V. Gozhyj, I. Kalinina, B. Rusyn, L. Pohreliuk, M. Korobchynskyi, Network security analysis based on consolidated threat resources, in: CEUR Workshop Proceedings, vol. 2604, 2020. URL: http://ceur-ws.org/Vol-2604/paper67.pdf.

[12] A. Geron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed., O'Reilly Media, 2019, 688 p.

[13] Artificial Intelligence: A Modern Approach. URL: https://towardsdatascience.com/understanding–the–bias–variance–tradeoff.

[14] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer-Verlag, 2009, 746 p.

[15] N. Purkait, Hands-On Neural Networks with Keras: Design and Create Neural Networks Using Deep Learning and Artificial Intelligence Principles, Packt Publishing, 2019, 462 p.

[16] E. Aydemir, C. Aktürk, M. A. Yalçınkaya, Estimation of housing prices with artificial intelligence, Turkish Studies 15 (2) (2020) 183–194.

[17] D. Sangani, K. Erickson, M. Al Hasan, Predicting zillow estimation error using linear regression and gradient boosting, in: Proceedings of the 2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), October 2017, pp. 530–534.

[18] C. Fan, Z. Cui, X. Zhong, House prices prediction with machine learning algorithms, in: Proceedings of the 2018 10th International Conference on Machine Learning and Computing, 2018, pp. 6–10.

[19] M. Yazdani, Machine learning, deep learning, and hedonic methods for real estate price prediction, arXiv preprint arXiv:2110.07151 (2021).

[20] J. Bin, S. Tang, Y. Liu, G. Wang, B. Gardiner, Z. Liu, E. Li, Regression model for appraisal of real estate using recurrent neural network and boosting tree, in: Proceedings of the 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), 2017, pp. 209–213.

[21] A. Varma, A. Sarma, S. Doshi, R. Nair, House price prediction using machine learning and neural networks, in: Proceedings of the 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 1936–1939.

[22] E. Walker, J. B. Birch, Influence measures in ridge regression, Technometrics 30 (2) (1988) 221–227.

[23] B. Afonso, L. Melo, W. Oliveira, S. Sousa, L. Berton, Housing prices prediction with a deep learning and random forest ensemble, in: Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional, 2019, pp. 389–400.

[24] A. K. Alexandridis, D. Karlis, D. Papastamos, D. Andritsos, Real estate valuation and forecasting in non-homogeneous markets: A case study in Greece during the financial crisis, Journal of the Operational Research Society 70 (10) (2019) 1769–1783. doi:10.1080/01605682.2018.1468864.

[25] J. L. Alfaro-Navarro, E. L. Cano, E. Alfaro-Cortés, N. García, M. Gámez, B. Larraz, A fully automated adjustment of ensemble methods in machine learning for modeling complex real estate systems, Complexity (2020) Article ID: 5287263. doi:10.1155/2020/5287263.

[26] A. Baldominos, I. Blanco, A. J. Moreno, R. Iturrarte, Ó. Bernárdez, C. Afonso, Identifying real estate opportunities using machine learning, Applied Sciences 8 (11) (2018) 2321. doi:10.3390/app8112321.

[27] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[28] Y. Liang, J. Wu, W. Wang, Y. Cao, B. Zhong, Z. Chen, Z. Li, Product marketing prediction based on XGBoost and LightGBM algorithm, in: Proceedings of the 2nd International Conference on Artificial Intelligence and Pattern Recognition, 2019, pp. 150–153.

[29] H. Seya, D. Shiroi, A comparison of residential apartment rent price predictions using a large data set: Kriging versus deep neural network, Geographical Analysis 54 (2) (2022) 239–260.

[30] J. G. Liu, X. L. Zhang, W. P. Wu, Application of fuzzy neural network for real estate prediction, in: International Symposium on Neural Networks, Springer, Berlin, Heidelberg, 2006, pp. 1187–1191.

[31] Dataset: Flats CSV. URL: https://www.kaggle.com/datasets/dmitruy11/flats-csv.

[32] I. Kalinina, A. Gozhyj, P. Bidyuk, V. Gozhyj, Multilevel ensemble approach in classification problems, in: Proceedings of the 2024 IEEE 19th International Conference on Computer Science and Information Technologies (CSIT), Lviv, Ukraine, 2024. doi:10.1109/CSIT65290.2024.10982625. URL: https://ieeexplore.ieee.org/document/10982625.

[33] I. Kalinina, A. Gozhyj, V. Gozhyi, S. Shiyan, Improving the architecture of a two-level heterogeneous ensemble for solving machine learning problems, in: Proceedings of the Intelligent Systems Workshop at the 9th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2025), Kharkiv, Ukraine, May 15–16, 2025. URL: https://ceur-ws.org/Vol-3983/.

[34] P. Bidyuk , I. Kalinina, O. Zhebko, A. Gozhyj, T. Hannichenko, Classification system based on ensemble methods for solving machine learning tasks, CEUR- WS. (2023), vol. 3426. Pp. 1-11. CEUR-WS.org/Vol-3426/paper5.pdf. (ISSN 1613-0073).

[35] G. Kunapuli, Ensemble Methods for Machine Learning, Manning, 2023, 352 p. ISBN: 9781638356707.

[36] G. Kyriakides, K. G. Margaritis, Hands-On Ensemble Learning with Python, Packt Publishing, 2019, 298 p. ISBN: 9781789617887.

[37] D. Mishra, Sh. M. Tripathi, A. Chaurasia, P. K. Chaurasia, A review on ensemble learning methods: Machine learning approach, International Journal of Research Publication and Reviews 6 (2) (2025) 3795–3803. doi:10.55248/gengpi.6.0225.0971.

[38] M. Shah, K. Gandhi, K. A. Patel, H. Kantawala, R. Patel, A. Kothari, Theoretical evaluation of ensemble machine learning techniques, in: Proceedings of the 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), 2023. doi:10.1109/ICSSIT55814.2023.10061139.

[39] N. A. Khleel, K. Nehéz, M. Fadulalla, A. Hisaen, Ensemble-based machine learning algorithms combined with near miss method for software bug prediction, International Journal of Networked and Distributed Computing (2025) 17 p. doi:10.1007/s44227-024-00044-x.

[40] A. A. Khan, O. Chaudhari, R. Chandra, A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation, Expert Systems with Applications 244 (2024) 122778. doi:10.1016/j.eswa.2023.122778.

[41] B. Naderalvojoud, T. Hernandez-Boussard, Improving machine learning with ensemble learning on observational healthcare data, AMIA Annual Symposium Proceedings (2023) 521–529. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10785929/.