Scientific Journal of Astana IT University
ISSN (P): 2707-9031
ISSN (E): 2707-904X

DOI: 10.37943/21EBBH5859

Saya Sapakova
Cand. of ph. and math. sc., Associate Professor, International University of Information Technology, Kazakhstan
s.sapakova@iitu.edu.kz, orcid.org/0000-0001-6541-6806
Zhansaya Bekaulova
Master of Technical Sciences, Assistant Professor, International Information Technology University, Kazakhstan
zh.bekaulova@iitu.edu.kz, orcid.org/0009-0000-9339-9222
Almas Nurlanuly
Master of technical science, Senior Lecturer, Civil aviation academy, Kazakhstan
a.nurlanuly@agakaz.kz, orcid.org/0000-0002-0364-0455
Duriya Daniyarova
PhD of Technical science, Associate Professor, International Educational corporation. Kazakh American University, Kazakhstan
duriya.daniyarova@mail.ru., orcid.org/0009-0000-5730-7407
Galiya Ybytayeva
PhD, Associate Professor, Department of Technical and Natural Sciences, International Educational Corporation, Kazakhstan
ybytayeva.galiya@gmail.com, orcid.org/0000-0002-4243-0928
Kaldybayeva Aizhan Seisebekovna
Senior Lecturer, Department of Information Technology and Librarianship, Kazakh National Women's Teacher Training University, Kazakhstan, aizhan.seisebek@gmail.com, orcid.org/0000-0002-2062-182X

DEVELOPMENT OF MACHINE LEARNING METHODS FOR MARKET TRENDS

Abstract: In the rapidly evolving real estate market, the application of machine learning (ML) is crucial for understanding and predicting price trends. This study evaluates and compares seven ML models, including multiple linear regression, random forest regression, support vector regression (SVR), decision tree regression, and XGBoost, to determine the most effective predictor of real estate prices in Astana, Kazakhstan. The study focuses on the Yesil district, a key area in the city, utilizing a dataset of over 9,000 records extracted from a broader collection of more than 30,000 real estate transactions across Kazakhstan. Through rigorous experimentation, model performance was assessed using statistical metrics such as mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (R-squared). The results indicate that the Random Forest Regressor and XGBRegressor models outperformed others, achieving the highest R-squared values (99.55% and 99.18%, respectively) and the lowest MAE and RMSE values. These findings highlight their robustness in predicting housing prices with high accuracy. The primary objective of this study was to develop a precise ML model capable of accurately forecasting real estate prices in Astana based on key market attributes. The superior predictive performance of the Random Forest and XGBRegressor models justifies their selection for deployment in real-world applications. Their high predictive accuracy suggests their potential utility for real estate professionals, policymakers, and investors seeking data-driven insights into market dynamics. This research expands of knowledge on the applications of ML in the real estate sector, reinforcing the importance of evidence-based decision-making within the industry.

Keywords*:* machine learning; real estate; data processing; regression analysis; algorithm.

Introduction

Within the contemporary landscape, the real estate market has garnered heightened scholarly attention, with a growing emphasis on leveraging sophisticated technological advancements, such as machine learning, to illuminate and anticipate its dynamic behavior. In this context, the integration of machine learning (ML) into the real estate sector presents a novel opportunity for enhanced market trend modeling. ML algorithms possess the capability to incorporate a multitude of factors, thereby facilitating the development of more efficacious property management strategies. Against this backdrop, this study focuses on the application of machine learning methods to analyze the online real estate market in Astana, offering an in-depth look at modeling and forecasting in this context. The choice of the Astana real estate market for the study was justified by its strategic location as the capital of Kazakhstan, the dynamic development of the city, changes in the geopolitical context, diversity of property types, investment potential and unique characteristics of the market in the Kazakh context

This research investigates the potential of employing artificial intelligence (AI) technologies for real estate market valuation [1]. The primary objectives are twofold:

- To assess the efficacy of various AI algorithms in determining real estate market value. This evaluation will be conducted by comparing the generated valuations with established accuracy benchmarks.
- To analyze the influence of specific algorithm parameters on the accuracy of valuation results. This will guide the selection of the most effective AI model for real estate market valuation within the chosen context.

Machine learning is revolutionizing real estate market research by providing accurate price forecasts, automated data analysis, identifying trends, identifying optimal investments, and classifying market segments. These technologies can also provide personalized recommendations, predict risks, and provide deeper, more informed insights into market dynamics. Real estate market research using machine learning is known to be used to optimize pricing, predict investments, provide personalized recommendations, make strategic decisions, manage risks, and improve customer experience. This data helps you make informed decisions and adapt to changing market conditions.

The real estate market is a set of organizationally oriented relations between subjects of the real estate market, in which, on the basis of market relations, transactions of purchase and sale, exchange, rent, rental or other transactions are carried out, as a result of which ownership rights or temporary procession of real estate are transferred for the purpose of exchanging existing rights to financial or other assets. We can observe that the real estate market in Kazakhstan has been fluctuating for several months. After recording high rates of real estate purchase and sale transactions, the demand curve went down, but prices per square meters continue to rise [2], [3].

The real estate market is influenced by many different factors from various fields and industries. It will be divided these factors into two types – external and internal. Studying external and internal factors in the real estate industry allows us to better understand market dynamics, predict changes and develop effective property management strategies. External factors include economic and geopolitical conditions and financial capabilities to the level of competition and diversity of property types. The interaction of these factors determines the dynamics of the real estate market, shapes prices and influences property management strategies.

For example, external factors include the military conflict in Ukraine, the policies of neighboring countries, which resulted in the depreciation of the national currency tenge. It should be noted here that these factors influenced the prices of building materials necessary for the construction of new residential buildings and also influenced the logistics infrastructure of our country. Internal factors include the termination of state mortgage programs, the possibility of withdrawing pension surpluses from the Unified National Pension Fund. Here it can be mentioned such state programs as “7-20-25”, “Baspana Hit”. These programs made mortgages more convenient but had many limitations. For example, the “7-20-25” program provided only for new housing, but the share of new housing in the regions is not as high as in the megacities of our country. This problem represents an important aspect of socio-economic dynamics, where the degree of housing affordability is closely related to the financial capabilities of the population. High property prices can put pressure on household budgets, while higher household incomes can create additional housing opportunities. The correlation between the median home price and median household income suggests that income level plays an important role in determining housing affordability. If the average household income is rising, this can alleviate the financial burden associated with buying a home, and vice versa. This has significant implications for the development of housing and financing policies and strategies aimed at increased people's access to housing [26], [27].

This study contributes by comparing various machine learning models, including traditional methods and advanced techniques like XGBoost, for real estate price prediction. It also applies essential preprocessing methods, such as feature scaling and handling missing data, often overlooked in similar research. Focused on the dynamic real estate market in Astana, with a dataset of over 30,000 entries, including 9,000 from the Yesil district, it offers valuable insights into the local market, distinguishing our work from existing studies.

Literature Review

In the existing literature, most of the research has involved the comparison of various methods that prove beneficial in predicting housing prices. The quantity and nature of attributes under consideration vary across these studies. The study by [2] utilizes a dataset of 5,359 townhouses in Fairfax County, Virginia, to investigate the application of machine learning algorithms for house price prediction. The research employs C4.5, RIPPER, Naïve Bayes, and AdaBoost algorithms to construct a predictive model. Subsequently, the study evaluates the classification accuracy of each model in forecasting house prices. The findings reveal that the RIPPER algorithm, known for its focus on high accuracy, consistently outperforms the other models in this task. In [3], an author presents a hybrid algorithm based on fuzzy linear regression (FLR) and fuzzy cognitive map (FCM) to solve the problem of forecasting and optimizing housing market fluctuations. The best fitting FLR model is then selected based on two metrics including confidence index (IC) and mean absolute percentage error (MAPE). To achieve this goal, analysis of variance (ANOVA) for a randomized complete block design (RCBD) is used. The proposed hybrid FLR-FCM algorithm allows decision makers to make use of imprecise and ambiguous data and more clearly represent the resulting model values. This is the first study to use a hybrid predictive approach to forecasting and optimizing housing prices and the market. The authors in [3] use the Random Forest machine learning method to predict housing prices and evaluate its effectiveness on a housing data set in the UCI Boston machine learning repository (507 records, 14 functions). The model shows acceptable predicted values, closely matching the actual prices within ± 5.

The paper [4] examines property price prediction models in Surabaya using Random Forest machine learning algorithms and seventeen regularly used characteristics from real estate agents that are the most influential factor in determining house prices. Another study on house price forecasting by Gierek [5] requires the use of the most accurate methodology to achieve maximum accuracy of the initial forecast. One of the methods that can be used to solve the problem of estimating the value of a house under uncertainty is fuzzy logic. Another study compares artificial neural networks with fuzzy logic and K-Nearest Neighbors to determine the most appropriate approach for pricing, which can be a useful guide for sellers [7],[8],[9]. Prediction of housing prices using a memristor-based artificial neural network was carried out by researchers Wang JJ et al [10]. To define a multivariate regression model using the backpropagation formula, they developed a synthetic neural network supported by memristors.

The author's work [11] examines the use of artificial intelligence, machine learning and nonlinear statistical models to solve problems of forecasting housing prices in Taiwan. In this order, the authors use ensemble regression boosting trees, support vector regression, and Gaussian process regression. Bayesian optimization is implemented through ten-fold cross-validation to determine the corresponding optimal kernels and parameter values. In the study described in [11], [12] the authors utilized the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm, which was trained with 28 variables selected through stepwise logistic regression to forecast housing prices in the United States [13], [14]. The RIPPER algorithm demonstrated superior performance compared to the C4.5 algorithm, Naïve Bayes, and AdaBoost algorithm. Additionally, in [15], [16] the researchers integrated Ensemble Empirical Mode Decomposition (EEMD) and Support Vector Regression to predict abrupt declines in house prices in the United States. The model presented in their work was trained using ten annual macroeconomic variables.

Other research has concentrated on utilizing decision trees to model and predict housing prices. For example, in [17], the random forest algorithm was employed to forecast the House Price Index in the United States, achieving a 5% error margin. Additionally, in the Australian market data context, decision trees, gradient boosting, and the random forest algorithm were found to be more effective compared to a multiple linear regression model [18]. Moreover, artificial neural networks demonstrated effectiveness in predicting house prices in various locations, including China (China Real Estate Index System) [19], Lagos (Nigeria) [20]. In a subsequent study [14], the authors found that the Kuala Lumpur house price forecasting model using XGBoost was highly effective, demonstrating the lowest MAE and RMSE values, as well as the best adjusted coefficient of determination (r-squared), consistently outperforming other machine learning models. In [21], [22] researchers used various machine learning methods, including random forest, gradient boosting, and XGBoost, to evaluate residential properties in California, Florida, and Texas. Comparative analysis with standard ordinary least squares (OLS) revealed superior performance of the machine learning models across all aspects. The model developed by the authors showed a median absolute percentage error of 9,3%. This increased accuracy coupled with cost-effectiveness and instant results, leads the authors of [22] to claim the superiority of automated scoring models over traditional methods. Zillow, a leading company in this area, as highlighted in [23], provides automated residential property appraisals with an astounding 3,5% accuracy (average error rate). Demonstrating its commitment to technological advancement, Zillow continually strives to improve the accuracy of its ratings. To achieve this goal, the company organized a competition [23] on the widely recognized Kaggle platform [24] between 2017 and 2018, offering a substantial prize fund of $1,200,000 for developing the best scoring algorithm [25].

Analysis of these studies shows that domestic valuation is in the early stages of introducing various artificial intelligence methods in the valuation of various assets [26]. Let’s look at the initial results of using machine learning methods to determine the value of residential real estate. This analysis is based on a thorough study of extensive data sets relating to apartment sales, using the example of apartment valuations on the secondary market in Astana.

Model Specification

A variety of machine learning techniques can be applied to assess the market value of real estate effectively. This section presents research findings on various machine learning approaches, emphasizing both traditional and contemporary methods. It discusses well-established techniques like linear regression, which is frequently utilized in estimation practice, alongside classic algorithms such as Random Forest and Gradient Boosting. Additionally, the study explores more recent and advanced models, including XGBoost, known for its efficiency and predictive accuracy.

This research specifically examines the following methods:

- Linear Regression
- Lasso Regression
- Random Forest Regressor
- Ridge Regression
- Support Vector Machine (SVM Regressor)
- Decision Tree
- XGBoost Regressor (XGBRegressor)

The selection of these methods is grounded in their adaptability to various data characteristics and robustness across different research scenarios. Notably, each method possesses unique features that enhance its applicability in real estate valuation. For instance, Lasso and Ridge regression techniques incorporate regularization to prevent overfitting, making them particularly useful in datasets with many features. Random Forest and Decision Tree models offer interpretability and handle non-linear relationships effectively. In contrast, XGBoost stands out due to its gradient boosting framework, which optimizes performance through sequential model training and regularization, resulting in superior predictive power. The uniqueness of this work lies in its comprehensive comparison of these diverse methodologies applied specifically to the dynamic real estate market of Astana, Kazakhstan. By synthesizing traditional and cutting-edge machine learning approaches, this study not only identifies the most effective predictors of real estate prices but also provides valuable insights into the local market characteristics. This contribution enhances decision-making processes for stakeholders and highlights the potential for tailored machine learning applications in real estate valuation.

The Support Vector Machines (SVM)

The Support Vector Machines (SVM) aims to find the optimal hyperplane (for binary classification) or separating hyperplane (for multiclass problems) by maximizing the margin between classes. The formula for linear SVM is expressed as follows:
For binary classification:

$$f(x) = w \cdot x + b,$$

where: $f(x)$ – decision function, w – weight vector, x – input data, b – bias term.
The objective is to find w and b and maximize the margin, subject to the constraint:

$$y_i \cdot (w \cdot x_i + b) \geq 1,$$

where:
y_i – class label (1 or -1),
x_i – training data.
SVM can also utilize kernels for handling nonlinear data. The formula for nonlinear SVM using kernel function $K\ (x, y)$ is:

$$f(x) = \sum_i^n a_i \cdot y_i \cdot K(x, x_i) + b, \quad (1)$$

where:
a_i – Lagrange multiplier,
n – number of training data points.
These formulas encapsulate the fundamental principles of SVM, aiming to find the optimal hyperplane or separating hyperplane using various kernels for nonlinear data.

Linear Regression

Linear Regression (LR) is a simple and widely used method for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship with the goal of finding the best-fitting line that maximizes the sum of squared differences between observed and predicted values.
The linear regression model can be represented as:

$$f(x) = w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + w_n \cdot \mathrm{x_n}, \quad (2)$$

here, $f(x)$ is the predicted value, w_0 is the intercept, $w_1, w_2, \ldots, w_n$ are the coefficients, $x_1, x_2, \ldots, x_n$ are the input features. The goal of LR is to find the coefficients of a linear equation that minimize the sum of squared differences between the predicted values and the actual values.

Lasso Regression

In Lasso Regression, and L1 regularization term is added to the linear regression objective function:

$$f(x) = w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + w_n \cdot x_n + \lambda \sum_{i=1}^{n} |w_i|. \tag{3}$$

here, $f(x)$ is the predicted value, $w_1, w_2, \ldots, w_n$ are the coefficients, $x_1, x_2, \ldots, x_n$ are the input features, λ is the regularization parameter. The additional term $\lambda \sum_{i=1}^{n} |w_i|$ penalizes the absolute values of the coefficients, encouraging sparsity in the model. The choice of λ determines the strength of the regularization.

Random Forest Regressor

Random Forest is an ensemble learning method that combines predictions from multiple decision trees to improve accuracy and reduce overfitting. While a single decision tree's formula is complex, the overall prediction in a random forest is the average (or median) of predictions from individual trees, enhancing robustness and predictive accuracy Figure 1.

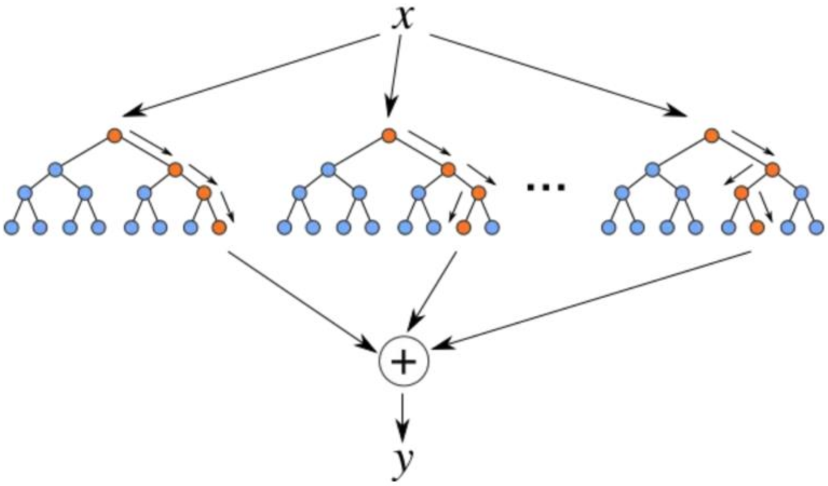


Figure 1. Random Forest Regressor architecture

Mathematically, if we denote the prediction of the i-th tree as $y_i(x)$ for input x, the Random Forest prediction $f(x)$ can be expressed as:

$$f(x) = \frac{1}{N} \sum_{i=1}^{n} y_i(x), \tag{4}$$

here, $f(x)$ is the overall prediction of the Random Forest for input *x*, *N* is the total number of trees in the forest, $y_i(x)$ is the prediction of the *i-th* tree for input *x*. Essentially, a random forest regressor takes multiple predictions generated by multiple trees and combines them through averaging, producing a prediction that is more consistent and accurate across regression scenarios.

Ridge Regression

Ridge Regression or L^2 regularization, is another variation of linear regression that introduces a penalty term based on the squared values of the regression coefficients. The objective function for Ridge Regression is as follows:

$$f(x) = w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + \cdots + w_n \cdot x_n + \lambda \sum_{i=1}^{n} |w_i^2|, \qquad (5)$$

here, λ is the regularization parameter that controls the strength of regularization. Ridge Regression helps prevent overfitting by penalizing large coefficients, providing a balance between simplicity and accuracy in the model.

Decision Tree

Tree structure: a decision tree is a hierarchical structure of nodes and leaves. Each node in the tree contains a condition, and each leaf contains a prediction.

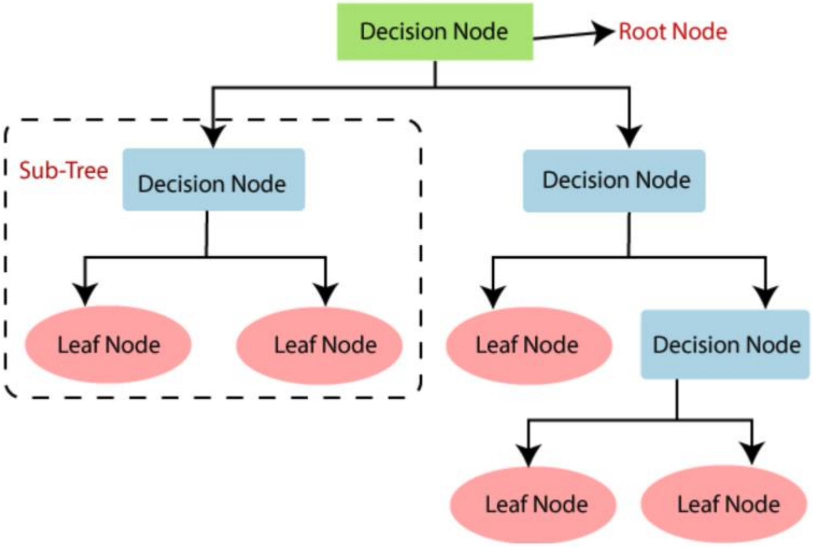


Figure 2. Decision tree architecture

At each level of the tree Fig.2, a feature and threshold are selected to split the data, minimizing uncertainty or error. This process is recursively repeated for each subgroup, creating new nodes. The splitting continues until a stopping criterion is met, at which point the nodes become leaves, each containing the final prediction.

XGBoost Regressor

Ensemble model: XGBoost (Extreme Gradient Boosting) is and ensemble of decision trees. Predictions from individual trees are combined to achieve a more accurate prediction. XGBoost is a powerful and scalable machine learning algorithm known for its efficiency and performance in structured/tabular data scenarios. It belongs to the class of gradient boosting algorithms and is widely used for regression, classification, and ranking tasks. XGBoost minimizes a loss function measuring the difference between predicted and actual values. The loss function includes an error component and regularization. Trees are added sequentially, and each tree corrects residual errors of previous ones. The training process stops when a certain criterion or a specified number of trees is reached. XGBoost incorporates regularization to prevent overfitting and enhance the model's generalization ability [11-12].

Prediction Formula:

$$\hat{y}_i(x) = \sum_{k=1}^{K} f(x)_k(x), \tag{6}$$

where:

$f(x)$ – the final prediction of the model,
$f(x)_k$ – prediction from an individual tree,
K – the total number of trees.

Both algorithms have their advantages and drawbacks, and the choice between them depends on the nature of the data and the requirements of the task.

Features of preparing initial data for training

Each property has a unique set of characteristics that affect its current market value. The goal is to formulate a rule based on data covering various properties. This data set includes the characteristic value of each property as well as the corresponding market prices. The goal is to establish a rule for predicting the likely market price of a new property given its characteristics. It is noteworthy that all factors influencing the market value are taken into account, including the technical characteristics of the property, its geographical location and the contextual market environment in which real estate transactions occur. The objectives of this study include analyzing the feasibility of determining the market value of real estate based on machine learning, evaluating the valuation results using generally accepted accuracy criteria for various methods (algorithms), selecting the most effective one, and studying the influence of the parameters of the selected algorithm on the accuracy of the results obtained.

The accuracy of property assessment is significantly influenced by the set of indicators, often referred to as pricing parameters, which are used to characterize the property under evaluation. Therefore, a crucial aspect of initializing the assessment process involves defining the composition of functions to describe each property. It is important to emphasize that the comprehensiveness of the property description is confined to the information available in sales advertisements published on relevant platforms. Through a comprehensive analysis of the real estate market, the following fundamental characteristics of a property, crucial in determining its market value, have been identified:

1) **Quantitative variables:**
- Year of construction;
- Number of floors and floor;
- Total area of the apartment;
- Kitchen area.

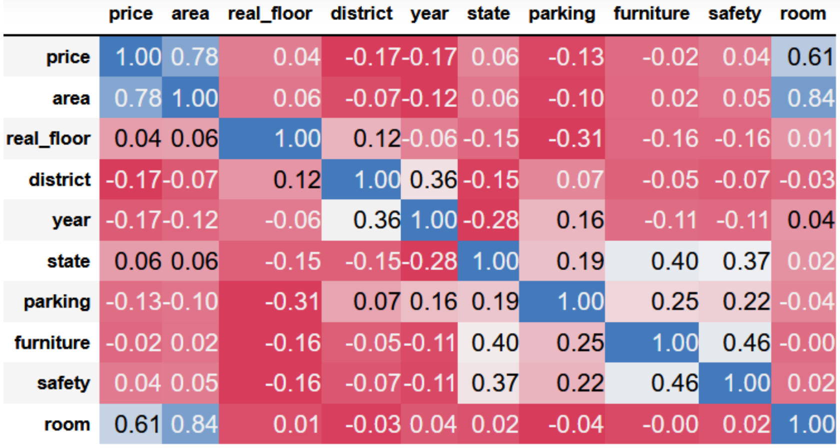
2) **Categorical variables:**
- District;
- Wall material;
- Furniture;
- Security features (safety);
- Perking;
- Availability of a balcony;
- Utilities/Communications;
- Balcony glazing (the presence of balcony glazing is of great importance, especially in connection with the harsh climatic conditions in our country);
- Infrastructure.

It is important to note that the specific set of parameters may vary depending on the source of price data. Typically, most of these characteristics are found in advertisements on various websites for the sale and rental of various real estate. However, not all advertisements provide comprehensive information about the properties for sale. The most complete information can often be found on a special website [25], which provides the necessary information about the location and condition of the apartment. However, even this site may not contain all important information. As for the location, this study adopted the practice of characterizing it by the price zone where the object is located. Zoning can be achieved through clustering according to various criteria, for example, object coordinates. Another problem that requires attention at the data preparation stage is related to the presence of various "noises" in advertisements. These may include outliers, false advertising with inflated or underpriced prices, and properties with inappropriate or conflicting features. To address this issue, the dataset chosen for the study was carefully prepared to exclude outliers and advertisements with inaccurate data, such as unrealistically low or high prices. When processing numerical characteristics, values falling below the 2nd percentile and above the 98th percentile were excluded. Additionally, to ensure consistency, all numeric attribute values have been standardized to a single type, such as integers or real numbers. This careful preparation aims to improve the quality and reliability of the data set for subsequent analysis. The table below shows the original characteristics obtained from the downloaded data site [25], and these descriptions are the same for all cities in Kazakhstan. However, it is obvious that not all data is complete: for some characteristics more than half of the values are missing. Additionally, most types of raw data are categorized as objects. Each field contained extraneous or mixed data. Before data preprocessing, we initially had 23 characteristics variable for each object. The price of the item serves as the dependent variable and target. The table provides significant information for explanatory purposes. For the study, more than 30 thousand data were collected from different regions of Kazakhstan, but in this work, we considered only the Yesil district of Astana. 9243 data were collected for this area.

To ensure compatibility with most models, categorical variables were converted into a numeric representation using the LabelEncoder method. As we know, LabelEncoder is a preprocessing technique in machine learning used to convert categorical (string) variables into numeric format. It assigns unique numeric labels to each unique value in a column, making it easier for machine learning models to handle categorical data. One of the most difficult tasks that an appraiser faces is structuring the information obtained from the ad text, especially from the Description characteristics (Table-1). The fact is that the most significant information about the value of the object being appraised is contained in the content of the ad, which is usually written in free form in accordance with the personal preferences of the author of the ad (seller, realtor, etc.). with proper processing, this description can be used to form the necessary features, and this improve the accuracy of model estimation. From this field you can select information about the location of the object, as well as information about the infrastructure. During the implementation of this study, various relevant features were identified. This made it possible to present a description of the object in a structured form and ensure further processing of this information. This article will give only one example of using a detailed text description to obtain a new featured.

Using the correlation matrix, we determined the degree of linear relationship between them. As is known, correlation coefficients vary from -1 to 1. As we see, the following factors turned out to be positive: `{('price', 'area'), ('price', 'real_floor'), ('price', 'state'), ('price', 'safety'), ('price', 'room')}`.

Using a correlation matrix-based heat map provides a quick and intuitive way to visualize the relationships between numerical variables in your data. As a results of the heat map, we got the following result:



	price	area	real_floor	district	year	state	parking	furniture	safety	room
price	1.00	0.78	0.04	-0.17	-0.17	0.06	-0.13	-0.02	0.04	0.61
area	0.78	1.00	0.06	-0.07	-0.12	0.06	-0.10	0.02	0.05	0.84
real_floor	0.04	0.06	1.00	0.12	-0.06	-0.15	-0.31	-0.16	-0.16	0.01
district	-0.17	-0.07	0.12	1.00	0.36	-0.15	0.07	-0.05	-0.07	-0.03
year	-0.17	-0.12	-0.06	0.36	1.00	-0.28	0.16	-0.11	-0.11	0.04
state	0.06	0.06	-0.15	-0.15	-0.28	1.00	0.19	0.40	0.37	0.02
parking	-0.13	-0.10	-0.31	0.07	0.16	0.19	1.00	0.25	0.22	-0.04
furniture	-0.02	0.02	-0.16	-0.05	-0.11	0.40	0.25	1.00	0.46	-0.00
safety	0.04	0.05	-0.16	-0.07	-0.11	0.37	0.22	0.46	1.00	0.02
room	0.61	0.84	0.01	-0.03	0.04	0.02	-0.04	-0.00	0.02	1.00

Figure 3. Correlation Heat Map

Analyzing the result of the heat map Fig.3, all factors with a low positive coefficient were removed and added a new characteristic: price per square meter, which is also calculated separately for each apartment. The target variable remains the price of real estate. Once again, a correlation analysis with the remaining factors are conducted, the resulting table looks like this (Figure 4).

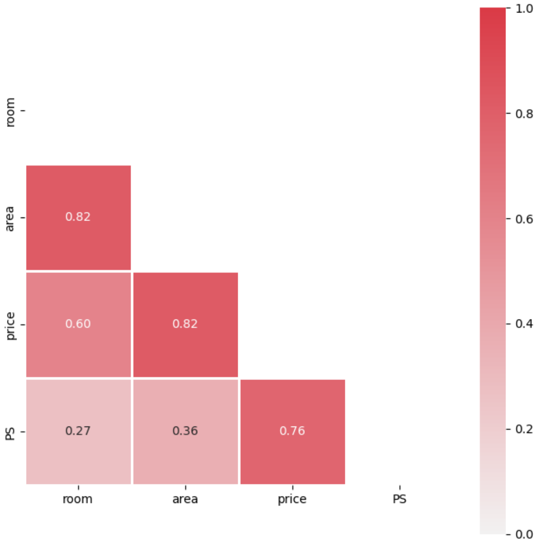


Figure 4. Heat Map of Pearson Correlation Coefficient Matrix

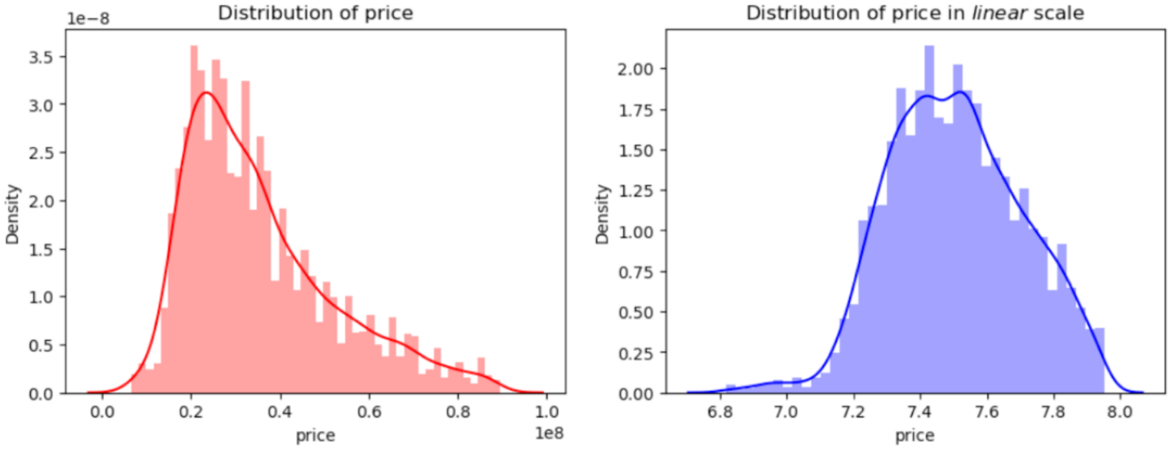


Figure 5. Distribution of price

The left subplot shows the original distribution (Figure 5), while the right subplot represents the distribution after applying the logarithmic transformation. The logarithmic transformation is often used to deal with skewed data and make it more suitable for certain types of analysis. After data preprocessing, initial computational experiments were carried out taking into account all factors. However, the algorithms demonstrated low accuracy in both training and testing, which is largely due to insignificant characteristics. In the second experiment, factors with negative correlation coefficients below 0.5 were excluded and a new factor PS was introduced. For the new data frame, all the above algorithms were carried out and their results were analyzed. The third stage of the experiment focused on time series, in particular the PS factor. To do this, an additional data set was collected for 2001, complete data on prices per square meter and exchange rates. Given Kazakhstan's strong correlation with the dollar exchange rate due to dependence on oil prices, an analysis was carried out of changes in the exchange rate over the past 23 years, as well as changes in prices per square meter. Economic dynamics made it possible to understand how changes in oil prices affect the strength of the national currency relative to foreign currencies. Skewness and kurtosis are known to play a key role in data mining, providing information about the shape of the distribution, the presence of outliers, and the need for transformation. These metrics help you understand how much data deviates from a normal distribution and how this can affect analysis and forecasting. This is important because many statistical methods assume normality of the data. Understanding the shape of the distribution helps you select appropriate models and data analysis methods, which can improve forecasting accuracy. Normal distribution of data facilitates the use of standard statistical methods such as t-tests and linear regression. To examine whether the dataset is normally distributed or not, the study used tei graphical methods such as histograms and boxplot.

Table 1. Kurtosis and skewness values for each column

Kurtosis values for each column		**Skewness values for each column**	
Before	**After**	**Before**	**After**
Room 0.663247 Area 21.564038 Price 93.964674 PS 163.491319	Room 0.663247 Area 0.598694 Price 0.463348 PS 0.060470	Room 0.654059 Area 3.320500 Price 7.421420 PS 7.833438	Room 0.285354 Area 0.895464 Price 0.999310 PS 0.281200

As we observe Table, there are no zero values in skewness, indicating that the original data is not symmetrical. The values for area (21.6), price (93.96), and PS (163.5) suggest positive skewness, indicating a longer right tail, as illustrated in the Figures 6-7. Analyzing kurtosis values, none are zero, indicating a non-normal distribution. The values for key characteristics room (0.65), area (3.32), price (7.42), PS (7.83) suggest a heavier tailed (sharper) distribution.

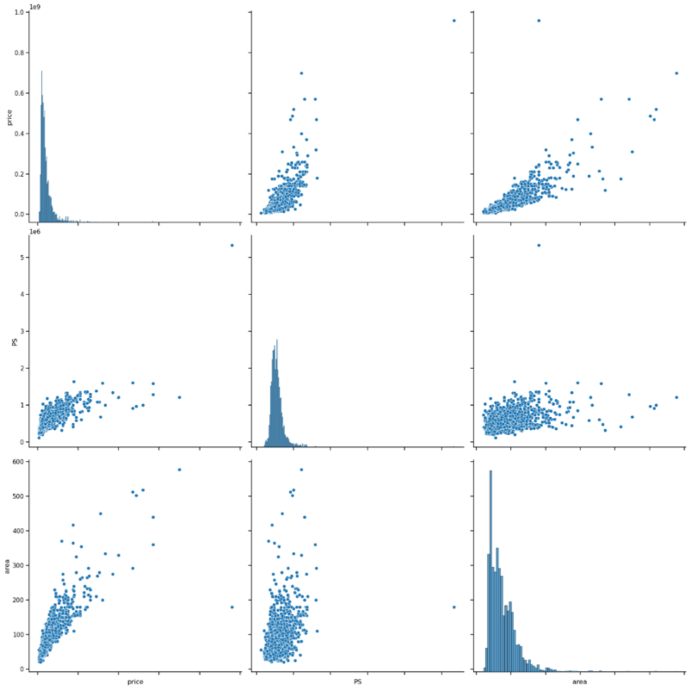


Figure 6. Pair plot Grid

Identifying outliers is critical when analyzing data because their presence can significantly affect the results. In our analysis, we used the interquartile range (IQR) method, which involves calculating the difference between the 25th percentile (Q1) and the 75th percentile (Q3) in a data set. This IQR, representing the spread of the central 50% values, was used to identify outliers. Specifically, an observation was considered an outlier if its value exceeded 1.5 times the IQR value or fell below 1.5 times the IQR. The modified dataset showed improved skewness and kurtosis values of {room: 0.285354; area: 0.8954; price: 0.999310; PS: 0.281200} and {room: 0.663247; area: 0.598694; price: 0.463348; PS: 0060470}, respectively. These numerical results are illustrated in Figures 1 and 2. A skewness values falling between -2 and 2, indicates a distribution that can be considered close to normal Fig. 8-9.

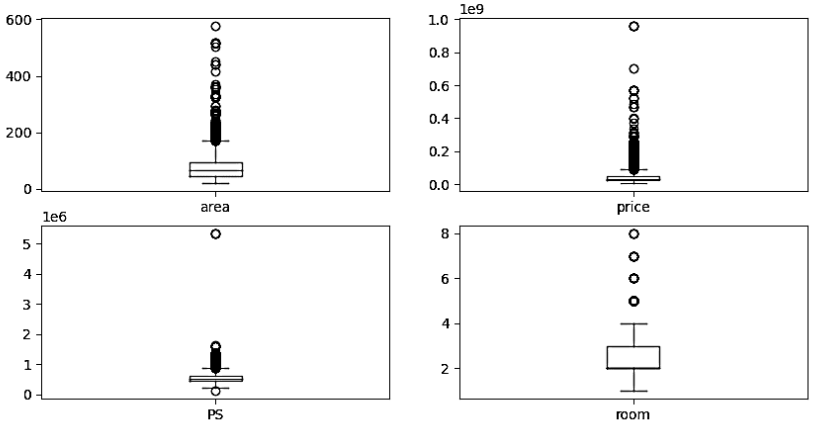


Figure 7. .Box plot result before

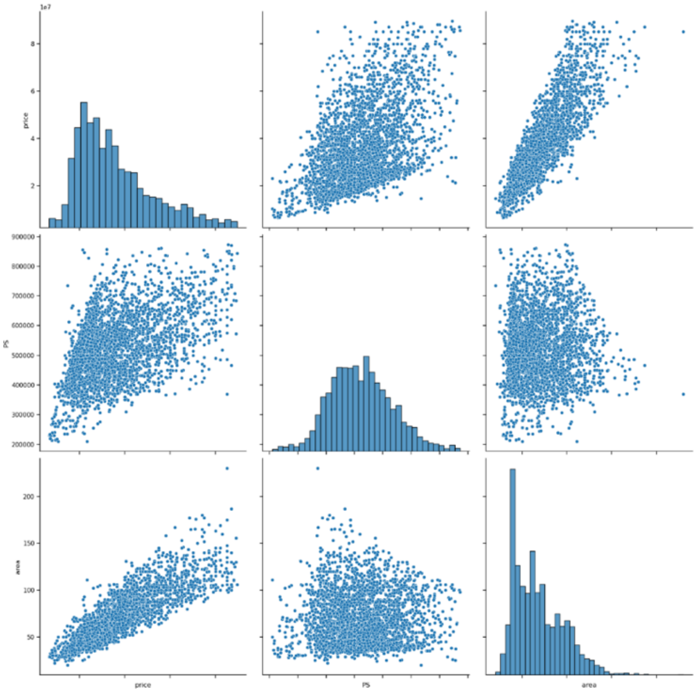


Figure 8. Pair plot after

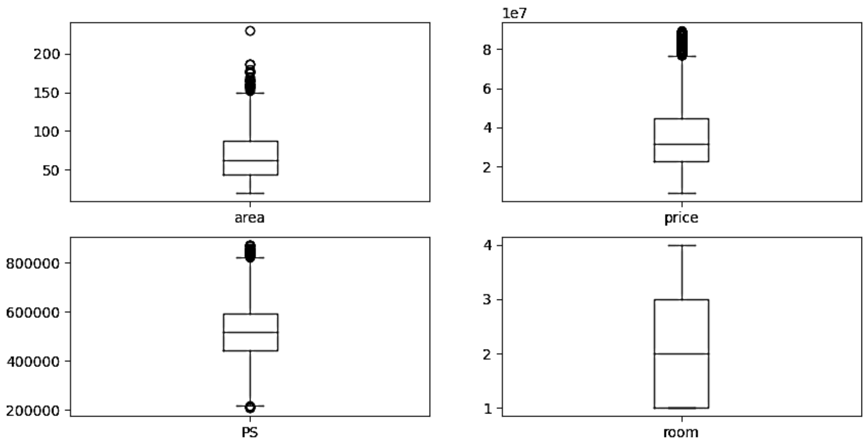


Figure 9. Box plot result after

5. Experiments and analysis of results

In this study, seven different models, including multiple linear regression, random forest regressor, SVM regressor, decision tree, and XGBRegressor, were implemented and compared to identify the most accurate predictor for real estate prices in the Yesil district of Astana. Performance metrics such as mean absolute error (MAE), root mean square error (RMSE), and R-squared were utilized to evaluate the models. The Random Forest Regressor and XGBRegressor models emerged as the top performers, showcasing the lowest MAE and RMSE values, along with an R-squared closest to unity, indicating an excellent fir compared to other considered models.

The primary objective was to establish a precise machine learning model capable of predicting real estate prices accurately. The study aimed to assess the relationship between the dependent variable (house price) and various independent variables (attributes). Statistical indicators, including MAE, RMSE, and R-squared, were employed for performance evaluation. The XGBRegressor and Random Forest Regressor models were selected for the deployment phase due to their superior predictive accuracy, as reflected in the highest coefficient of determination (R-squared). This choice signifies that these models are the most suitable for accurately predicting housing prices in Astana based on the provided dataset.

Table 2. Comparison of algorithm results

№	Model	MAE	MSE	RMSE	R2 Square
0	Linear Regression	7,3897	1,6072	1,2678	0,9915
1	Lasso Regression	8,1366	1,8459	1,3586	0,8983
2	Random Forest Regressor	0,5929	0,0821	0,2865	0,9955
3	Ridge Regressor	7,6814	1,6605	1,2886	0,9085
4	SVM Regressor	15,6653	10,526	3,2445	0,4203
5	Decision Tree	1,0330	0,5206	9,7215	0,9713
6	XGBRegressor	0,8072	0,1481	0,3848	0,9918

The Random Forest Regressor model exhibits an exceptionally high R-squared value of 0.9955, indicating an outstanding fit to the data Table 2. This suggests that the model explains a substantial portion of the variance in the target variable and performs exceptionally well in capturing the underlying patterns in the dataset. Additionally, A cross-validation score of 0.91889 indicates a high generalization capability of the model. This suggests that the model is likely to perform well on new, unseen data, and its results on test data can be considered reliable. In general, the closer the cross-validation score to 1, the better the model’s generalization. Overall, these results suggest that the random forest model is a good fit for the data and can be used for making accurate predictions on new data. The SVM regressor shows moderate explanatory power (R2: 0.42) but raises concerns about generalization performance, as indicated by the low cross-validation score (0.000). in conclusion, based on the provided metrics, the Random Forest regressor outperforms the SVM Regressor in terms of explanatory power and overall performance.

Conclusion

This study delves into the complexity of estimating residential property values using intrinsic characteristics using a wide range of machine learning and data mining techniques. The study evaluates the performance of the model using standard metrics MAE, MSE, RMSE and R^2. A comparative analysis of these indicators was carried out based on data obtained from real estate websites in Kazakhstan. Seven machine learning algorithms were trained, and the inclusion of additional features improved the overall qualities of the models Table. Notably, the random forest regressor algorithm demonstrated excellent performance, achieving an R2 accuracy of 0.99 on the test dataset. These results confirm the effectiveness of machine learning in real estate appraisals and suggest practical applications for improving the quality of reports in appraisal firms. Moreover, the methods used can be extended to the pricing of commercial real estate and optimization of mass valuation of land plots. Future research could focus on further improving accuracy by exploring features such as identifying renovation types from detailed listing descriptions, using larger training datasets, and exploring the potential of convolutional neural networks for categorizing apartment renovations using visual data extracted from images.

References

[1] Pengyu, Chen. (2024). Using Machine Learning to Predict the Stock Market Trend. doi: 10.62051/bgrbnh39

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

[3] Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data [Expert Systems with Applications, 42 (2015), 2928–2934]. *Expert Systems with Applications*, *42*(19), 6806.

[4] Azadeh, A., Ziaei, B., & Moghaddam, M. (2012). A hybrid fuzzy regression-fuzzy cognitive map algorithm for forecasting and optimization of housing market fluctuations. *Expert Systems with Applications*, 39(1), 298–315.

[5] Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House price prediction using random forest machine learning technique. *Procedia Computer Science*, *199*, 806–813.

[6] TEM Journal. (2023). House price prediction model using random forest in Surabaya City. *TEM Journal, 12(1)*, 126–132. https://doi.org/10.18421/TEM121-17

[7] Gerek, I. H. (2014). House selling price assessment using two different adaptive neuro-fuzzy techniques. Automation in Construction, 41, 33–39.

[8] Manasa, N., David, Winster, Praveenraj., Lakshmi., S.R. (2023). Predictive Analytics for Stock Market Trends using Machine Learning. 1-8. doi: 10.1109/iccakm58659.2023.10449528

[9] Mukhlishin, M. F., Saputra, R., & Wibowo, A. (2017, November). Predicting house sale price using fuzzy logic, artificial neural network and K-nearest neighbor. In *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)* (pp. 171–176). IEEE. https://doi.org/10.1109/ICICOS.2017.8276357

[10] Phan, T. D. (2018, December). Housing price prediction using machine learning algorithms: The case of Melbourne city, Australia. In *2018 International Conference on Machine Learning and Data Engineering (iCMLDE* (pp. 35–42). IEEE. https://doi.org/10.1109/iCMLDE.2018.00017

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

[12] Wang, J. J., Hu, S. G., Zhan, X. T., Luo, Q., Yu, Q., Liu, Z., ... & Liu, Y. (2018). Predicting house price with a memristor-based artificial neural network. IEEE Access, 6, 16523–16528. https://doi.org/10.1109/ACCESS.2018.2814065

[13] Lahmiri, S., Bekiros, S., & Avdoulas, C. (2023). A comparative assessment of machine learning methods for predicting housing prices using Bayesian optimization. Decision Analytics Journal, 6, 100166.

[14] Wang, X., Wen, J., Zhang, Y., & Wang, Y. (2014). Real estate price forecasting based on SVM optimized by PSO. *Optik, 125*, 1439–1443.

[15]. Plakandaras, V., Gupta, R., Gogas, P., & Papadimitriou, T. (2015). Forecasting the U.S. real house price index. *Economic Modelling, 45*, 259–267.

[16] Yeh, I.-C., & Hsu, T.-K. (2018). Building real estate valuation models with a comparative approach through case-based reasoning. *Applied Soft Computing*, *65*, 260–271.

[17] Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2022). House price prediction using random forest machine learning technique. *Procedia Computer Science*, *199*, 806–813.

[18] Soltani, A., Heydari, M., Aghaei, F., & Pettit, C. J. (2022). Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms. Cities, 131, Article 103941.

[19] Xu, X., & Zhang, Y. (2021). House price forecasting with neural networks. *Intelligent Systems with Applications*, *12*, Article 200052.

[20] Lahmiri, S. (2018). Minute-ahead stock price forecasting based on singular spectrum analysis and support vector regression. *Applied Mathematics and Computation*, *320*, 444–451.

[21] Lahmiri, S., & Bekiros, S. (2020). Intelligent forecasting with machine learning trading systems in chaotic intraday Bitcoin market. *Chaos, Solitons & Fractals, 133*.

[22] Abdul-Rahman, S., Zulkifley, N. H., Ibrahim, I., & Mutalib, S. (2021). Advanced machine learning algorithms for house price prediction: Case study in Kuala Lumpur. *International Journal of Advanced Computer Science and Applications*, *12*(12). https://doi.org/10.14569/IJACSA.2021.0121291

[23] Kok, N., Koponen, E.-L., & Martinez-Barbosa, C. A. (2017). Big data in real estate: From manual appraisal to automated valuation. *The Journal of Portfolio Management*, *43*(6), 202–211.

[24] Zillow Prize. (n.d.). Zillow's home value prediction (Zestimate) [Website]. Retrieved from https://www.kaggle.com

[25] Kaggle. (n.d.). Your home for data science [Website]. Retrieved from https://www.kaggle.com

[26] Forbes.kz. (n.d.). *When will the 7-20-25 and Baspana Hit programs end in Kazakhstan?* Forbes Kazakhstan. https://forbes.kz/articles/kogda_v_kazahstane_zavershatsya_programmyi_7-20-25_i_baspana_hit

[27] krisha.kz. URL: https://www.krisha.kz.