IRSTI 06.52.35 https://doi.org/10.26577/jpcsit2025338

Ali Rakhimzhanov[1*], Jelena.V. Caiko [2]

[1]Kazakh-British Technical University, Almaty, Kazakhstan
[2]Riga Technical University, Riga, Latvia
*e-mail: ali.rakhimzhanov11@gmail.com

FORECAST OF HOUSING PRICES IN ALMATY USING MACHINE LEARNING ALGORITHMS

Abstract. Precise prediction of housing values is an important task for various stakeholders involved in the housing market, including investors, builders, and city planners. In this research, supervised machine learning models are used to predict the price of apartments in Almaty, Kazakhstan, which is a dynamic urban market in Central Asia. With an openly available dataset of apartments for sale, Linear Regression, Lasso Regression, Random Forest, and XGBoost models are implemented and tested. The data is scaled and encoded with scalable pipelines, and models are evaluated with regards to Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score. The best performing model amongst those tested was Random Forest Regressor with an R^2 of 0.9158, followed by XGBoost with 0.8438. Feature importance visualization identifies district, area, and construction year as primary influencing factors. The research supports that ensembling machine learning models are efficient and scalable predictors for housing forecasts and suggests future improvements with time-series and geospatial features.

Keywords: Housing price prediction; Machine learning algorithms; Regression models; Random Forest; Real estate market analysis; Urban economy.

1. Introduction

Housing markets worldwide have witnessed unprecedented deviations through economic, social, and political drivers in recent years. Almaty, Kazakhstan, and urban hubs of emerging economies around the world, have not escaped these trends. As the largest city and economic capital of the country, Almaty, with its high urbanization growth rate, has witnessed heightened demand for housing, infrastructural growth, and an uptrend for housing prices. Official numbers have Almaty place consistently at or near the head of the list of highest property prices per square meter in Kazakhstan[1], but pricing patterns remain geographically localized and data-dispersed. Such elements inject uncertainty and inconsistency into housing price valuation, posing a most sensitive challenge to stakeholders along the property spectrum–including buyers and sellers, as well as developers, banks, and policymakers.

Traditional housing price appraisal methods depend on statistical methods or human judgment, both of which suffer from inherent drawbacks. Linear regression models, for example, are unable to identify nonlinear relationships and interactions between features that exist in housing data. Further, these approaches are prone to human bias and tend to fail to cope with fast-evolving market conditions. On the other hand, machine learning (ML) algorithms have proven to be high-performance tools with the capability to learn intricate patterns based on large databases and make precise forecasts. The fact that ML models can perform prediction automatically and improve with time makes them the best fit for real estate prediction [3][2].

Around the world, several studies have established machine learning's capability for accurate price prediction of real estate. Decision Trees, Random Forest, Gradient Boosting Machines (i.e., XGBoost), and Artificial Neural Networks have been proven to outshine traditional methods, particularly for large, heterogenic datasets. For instance, studies done in China [4], India[5], and the United States [6] have established ML's potential for enhancing house price estimation. The majority of these studies, albeit, are based on developed real estate markets, where data access and quality facilitate such analyses. Conversely, there is a clear void of scholarly literature focusing on Central Asia and, more specifically, Kazakhstan, where there is underutilization of real estate data for model

analyses. It is this gap that creates an imperative for regionally based studies utilizing contemporary data science approaches for emerging markets.

This paper seeks to close that gap by examining machine learning algorithm applicability and effectiveness for residential housing price prediction in Almaty. We gather and preprocess actual estate listings from open data sources, extract features that are applicable, and use various regression-based ML models to fit and validate price forecasts. Models considered include Linear Regression, Lasso Regression, Random Forest, and XGBoost. Model performance was evaluated using three widely accepted metrics: the coefficient of determination (R^2, Eq. 1), Root Mean Squared Error (RMSE, Eq. 2), and Mean Absolute Error (MAE, Eq. 3). These metrics provide complementary perspectives on predictive accuracy, penalizing systematic bias, large deviations, and overall fit respectively.

The contributions of this study are three: First, we propose an extensive ML-based framework specific to Almaty's housing market, based on available public datasets and sophisticated algorithms. Second, we compare several different models to arrive at a suitable method for housing price prediction under conditions specific to Almaty. Third, we make actionable contributions to understanding what drives housing price dynamics in Almaty, which can help stakeholders make informed decisions based on data.

The rest of this paper is outlined below: Section 2 contains a literature review of machine learning and housing price forecasting. Section 3 outlines the data, feature engineering, and data preprocessing steps. Section 4 states the methodology, including model and training procedures. Section 5 presents results and evaluation. Section 6 concludes and offers future research guidelines and recommendations.

2. Literature review

Increases in data science adoption within real estate analytics have resulted in dramatic improvements to housing price forecasting. Historically, valuation was controlled by hedonic methodologies, which employed linear regression to link attributes of housing (size, location, number of bedrooms, etc.) to price. Though successful under restricted circumstances, these models fail to represent nonlinearities and high-level interactions between features, which are common under dynamic housing conditions. As a result, ML methods have become favored because of high-dimensional modeling, automatic discovery of patterns, and improved predictive performance under various market conditions.

The theoretical foundation of these approaches lies in the hedonic pricing model [7], which has long been used to capture structural and locational effects in property values. However, subsequent research demonstrated that linear frameworks struggle with multicollinearity and fail to incorporate spatial dependencies effectively [8].

Among the first to criticize traditional approaches was [2] who presented the shortcomings of a hedonic price model and promoted more adaptive, data-driven methods, particularly for differentiated urban contexts. A later review by [3] reinforced this development, illustrating that Random Forest, XGBoost, and Artificial Neural Networks outperformed linear models substantially through international case studies.

Other early explorations of machine learning in housing price prediction, such as [9], also demonstrated the potential of structured data integration (e.g., energy efficiency, accessibility), showing that nontraditional features could strongly influence property valuations.

Specifically, [4] proposed an XGBoost-driven housing price prediction model for urban China. It identified that the algorithm performed well in dealing with both high-dimensional inputs and missing data and surpassed Decision Trees and Ridge Regression with better RMSE and R^2 scores. Analogously, [5] surveyed more than a hundred research papers and concluded that ensemble learning methods, namely Random Forest and boosting methods, generalized well on actual housing datasets. Notably, Random Forest achieved the lowest RMSE (Eq. 2), confirming its robustness across error measures

Researchers have pursued hybrid methods to increase prediction accuracy. As an example, [10] employed a stacked model with Linear Regression, Random Forest, and XGBoost to achieve high precision for a Turkish housing market dataset. Another creative solution was proposed by [11] who integrated their prediction model into an MLOps pipeline, which supports auto-deployment and perpetual optimization within live real estate platforms.

In spite of this quick progress, research in emerging markets, and especially Central Asia, is

still limited. A major contribution to this field is provided by [12], who implemented Naive Bayes, Decision Trees, and AdaBoost on housing market data in Kazakhstan. Their research shows that machine learning can identify insightful market patterns and price drivers with only limited data available locally. This regional lack of research emphasizes how crucial and innovative it is to have localized ML models for cities such as Almaty.

In addition, deep learning–based models have been tested in international contexts (Cheng & Wang, 2018) [13], though their higher computational demand and sensitivity to feature scaling make them less commonly adopted in smaller-scale or emerging market studies compared to ensemble methods.

More recent studies are moving towards spatiotemporal modeling as well. [14] investigated applying geographically and temporally weighted machine learning, and demonstrated that including neighborhood-level and seasonal dynamics greatly enhances Sydney's forecasting. In a similar vein, [15] used Explainable AI (XAI) methods to model affordable housing dynamics with land value and zoning data, with a focus on interpretability for urban planning.

Recent research also considered feature engineering and multiobjective optimization. [16] employed evolutionary algorithms to hybridize ML with optimization methodologies and demonstrated that models with domain-specific objectives perform well in actual deployments. The relevance of hyperparameter fine-tuning, cross-validation, and importance of features has been reinforced through most recent literature, and ensembling models have remained at or near the top of performance and stability rankings. Furthermore, [17] suggested a hybrid model of TLBO and XGBoost with inherent uncertainty estimations to provide confidence-scored predictions for construction and real estate evaluations. Further, research as presented by [18][19] identifies the trend towards universal AI frameworks for construction and real estate.

3. Data Description and Preprocessing

3.1 Dataset Overview

We obtained data for this research through Kaggle [20], and it contains 11,883 residential apartments for sale in Almaty, Kazakhstan. Each record is a unique listing and includes rich features that specify the physical attributes, address, and price of the selling apartment. The dataset contains a variety of different types of apartments, including those found within Soviet-era structures to those newly developed high-rise buildings.

Table 1 – Features for Modeling

Feature Name	Original Data Type	Description / Unit	Role
price	Numerical	Total apartment sale price (in KZT)	Target variable
area	Numerical	Apartment floor area (in square meters)	Input feature
no_of_rooms	Numerical	Number of rooms	Input feature
floor	Numerical	Floor level	Input feature
year_of_construction	Numerical	Year building was constructed	Input feature
district	Categorical	Administrative district of Almaty	Input feature
structure_type	Categorical	Type of building construction (e.g., Brick)	Input feature
quality	Ordinal	Subjective quality score (Very Poor to Excellent)	Input feature
Id	Nominal	Unique listing identifier	Dropped
price_per_sqm	Derived (Numerical)	Price divided by area (KZT per m²)	EDA-only

3.2 Preprocessing

3.2.1 Data Cleaning

During initial data inspection, there were no missing or null values found in important fields. Duplicates were checked based on the Id field and deleted as required. All the numeric fields,price, area, and year of construction,were within desirable limits, reflecting good data consistency.

3.2.2 Feature Decoding

The original data had several categorical features encoded numerically. These were interpreted as follows:

- districts: 0–7 → Almalinsky, Auezovsky, Bostandyk, etc.
- structure_type: 0–3 → Panel, Brick, Monolithic, Other
- quality: 0–4 → Very Poor to Excellent

3.2.3 Feature Engineering

A derived variable, price per square meter, was defined by the formula:

price_per_sqm = price / area

The engineered feature price_per_sqm was calculated only for exploratory analysis purposes and was not included in the set of input features used for model training or testing.

Using this variable as a predictor would create target leakage, since it is a transformation of the target (price).

Instead, we used price_per_sqm only for visualizations in EDA (e.g., price distribution plots by district) to help understand pricing patterns and variability.

3.2.4 Summary Statistics

As indicated by Table 2, Almaty's average apartment costs around 55.7 million KZT and measures around 67.9 square meters. Price and area, as expected, both show wide variation, reflecting a highly diverse market. Construction years span from 1932 to 2023, showing both old Soviet structures and new developments.

Table 2 – Descriptive Statistics of Main Apartment Features in Almaty

Feature	Mean	Std. Dev.	Min	25%	Median	75%	Max
Price (KZT)	55,745,160	507,514,200	3,500,000	27,900,000	35,000,000	47,000,000	5.7B+
Area (m²)	67.9	42.9	8.0	44.0	59.5	80.0	600+
Rooms	2.24	1.01	1	1	2	3	5+
Floor	5.13	4.11	1	2	5	7	24
Year Built	2001	20.7	1932	1982	2008	2020	2023

3.3 Statistical Tests and Feature Diagnostics

• Normality Tests:

In addition to visual inspection of price and area histograms, we conducted statistical tests to examine distributional assumptions. Both Shapiro–Wilk and Kolmogorov–Smirnov tests rejected the null hypothesis of normality for apartment prices ($p < 0.001$) and living area ($p < 0.01$), confirming skewness and heavy tails.

• Multicollinearity:

Variance Inflation Factor (VIF) analysis indicated no severe multicollinearity (VIF < 5 across predictors). However, strong correlations were found between area, number of rooms, and price, which may explain instability in linear models.

• Feature Importance Beyond Gain:

SHAP (SHapley Additive exPlanations) values were used to interpret feature contributions beyond the XGBoost gain metric. SHAP confirmed that district, area, and quality were dominant predictors, but also revealed nonlinear effects (e.g., diminishing returns for very large apartments).

3.3 Exploratory Visualizations

To better understand data distribution and detect potential modeling issues, the following visualizations were created:

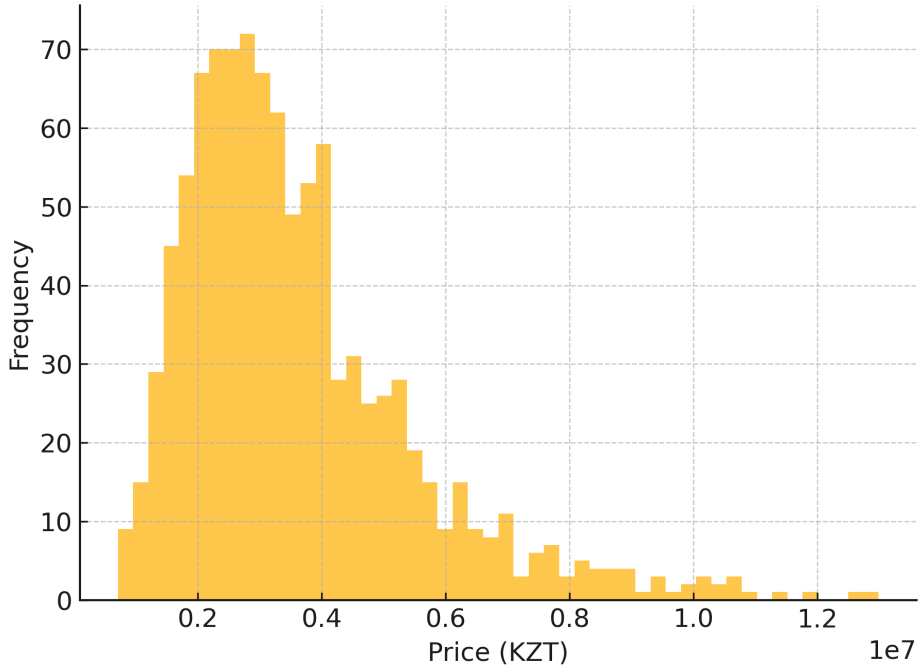

Figure 1 – Distribution of apartment prices in Almaty.

Price Distribution: Skewed to the right; most apartments are priced between 10 and 50 million KZT. Some listings exceed 500 million KZT, creating a long tail,indicating luxury segments that modeling needs to account for.

Area Distribution: Apartments range from compact 8 m² units to over 200 m². Most listings fall between 40–70 m², aligning with typical urban layouts.

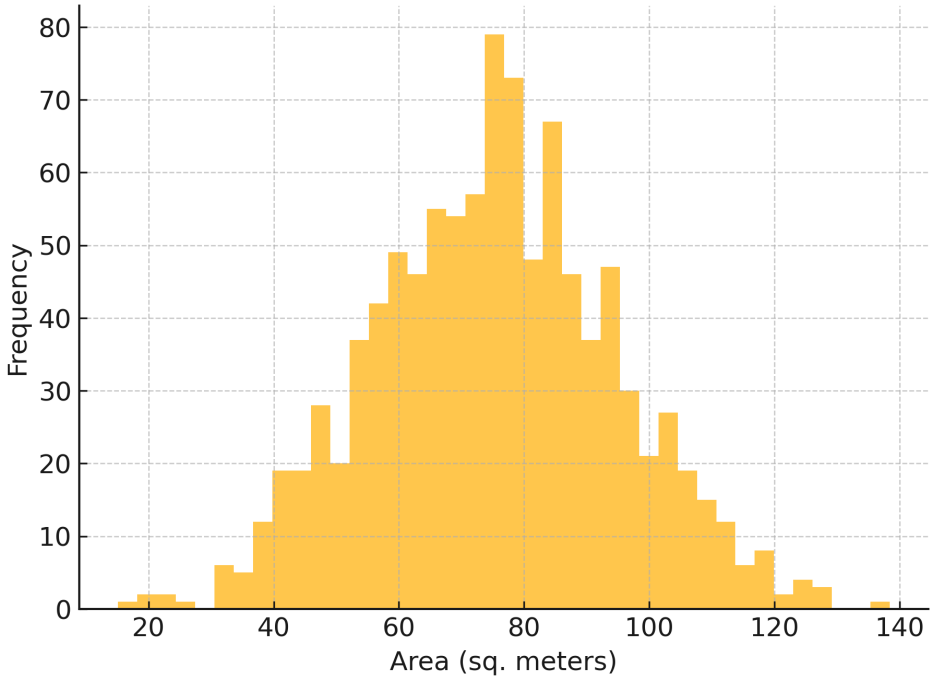

Figure 2 – Distribution of apartment areas (square meters).

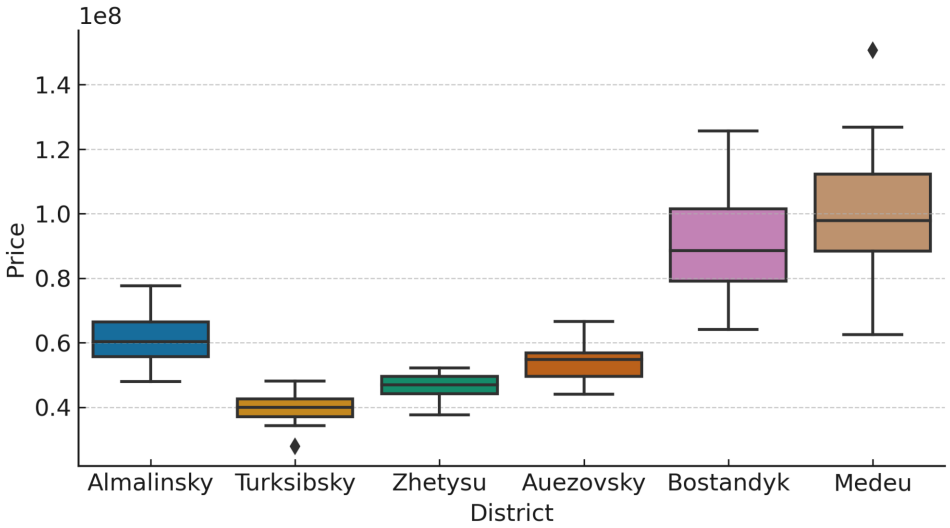

Figure 3 – Relationship between apartment area and price.

District Price Boxplots: Clear segmentation exists. Districts like Medeu and Bostandyk have higher median prices due to prestige and centrality. In contrast, Turksibsky and Zhetysu show lower median and range values. The width of price distribution also varies by district,suggesting differing volatilities.

3.4 Data Splitting and Scaling

For modeling purposes, the dataset was divided as follows:

- Training set: 80%
- Test set: 20%

A fixed random seed ensures reproducibility. For models sensitive to the magnitude of input features (like Support Vector Regression or KNN), Min-Max Scaling or Standardization will be applied as needed.

4. Methodology

4.1 Problem Formulation

This research is focused on forecasting apartment sale prices in Almaty using supervised machine learning. The challenge is treated as a regression task, where the inputs are structured housing attributes, and the output is the price prediction in Kazakhstani Tenge (KZT)

Let:

- $\boldsymbol{X} = [x_1, x_2, ..., x_n]$ be the set of input features (e.g., area, district, floor)
- $y \in \mathbb{R}$ be the apartment's sale price
- The model aims to learn a function $f : \boldsymbol{X} \rightarrow y$ that minimizes prediction error

4.2 Model Selection

In this study, we evaluated four supervised learning models commonly applied in housing price prediction: Linear Regression, Lasso Regression, Random Forest, and XGBoost. These were selected to balance interpretability, predictive power, and computational feasibility given the dataset size and regional context.

- Linear Regression was included as a baseline model, reflecting its long-standing use in traditional hedonic pricing frameworks, where housing attributes such as size, rooms, and location are linearly associated with price [2]
- Lasso Regression extends this baseline by incorporating L1 regularization, which reduces overfitting and can automatically perform feature selection by shrinking non-informative coefficients to zero [3]
- Random Forest is a robust, non-parametric ensemble method capable of capturing nonlinear feature interactions. It has consistently delivered strong performance in housing price prediction tasks across diverse markets, especially when datasets include both categorical and numerical variables [5]
- XGBoost, a gradient boosting framework, iteratively reduces residual errors and integrates regularization, enabling superior performance on heterogeneous datasets. Prior research has shown that boosting models often outperform bagging approaches in real estate forecasting [4]

We did not include other models such as Support Vector Machines (SVM) or Neural Networks because of their higher sensitivity to feature scaling,

risk of overfitting on tabular datasets, and the significant computational cost of tuning. Similarly, although LightGBM is considered a strong competitor to XGBoost, it was not evaluated here due to resource limitations. These alternatives remain promising directions for future research.

Overall, the selected models reflect a balance between theoretical grounding in the hedonic pricing tradition and the demonstrated success of ensemble learners in tabular prediction tasks. Prior comparative studies confirm that tree-based ensembles consistently outperform kernel-based methods such as SVM in structured housing datasets, while maintaining lower computational overhead than deep learning models. In this context, the four chosen models represent both methodological diversity and practical feasibility for the Almaty housing market.

4.3 Pipeline and Preprocessing

All models were implemented using scikit-learn and XGBoost, with a reproducible pipeline design to ensure consistency. The preprocessing steps included:

• Numerical Features: Scaled using StandardScaler.

• Categorical Features: One-hot encoded using OneHotEncoder(drop='first').

• Train/Test Split: The dataset was split into 80% training and 20% testing using a fixed random seed for reproducibility.

Outlier Removal: Outliers were detected based on values exceeding three standard deviations from the mean in either price or area. Approximately 2.7% of the dataset (322 listings out of 11,883) were removed. To avoid data leakage, the mean and standard deviation were computed only from the training set, and the same thresholds were then applied to filter the test set.

This step improved model stability, particularly for ensemble methods, which are sensitive to extreme target values. Linear models were less affected, but overall performance was enhanced by excluding unrealistic outliers (e.g., one apartment listed at over 5.7B KZT).

4.4 Evaluation Metrics

Three standard regression metrics were used to assess model performance:

1. Mean Absolute Error (MAE) Measures average error size without regard to direction. $MAE = (1/n) \times \sum |y_i - \hat{y}_i|$

2. Root Mean Squared Error (RMSE) Heavily penalizes larger errors. $RMSE = \sqrt{[(1/n) \times \sum (y_i - \hat{y}_i)^2]}$

3. R^2 Score (Coefficient of Determination) Shows how much of the variance in target prices is explained by the model

$$R^2 = 1 - (\sum (y_i - \hat{y}_i)^2 / \sum (y_i - \bar{y})^2)$$

4.5 Implementation Details

All modeling work was conducted in Python 3.10 using the scikit-learn and xgboost libraries.

A 5-fold cross-validation scheme was consistently applied across all models. For Linear Regression and Lasso Regression, CV was used to check performance stability and ensure robustness. For Random Forest and XGBoost, CV was also integrated into the hyperparameter search. Final results reported in Section 5 are based on the held-out test set (20%).

Hyperparameter Tuning**:** A limited grid search was conducted for the Random Forest model with n_estimators ∈ {100, 200, 300} and max_depth ∈ {5, 10, 15}. The configuration that provided the most stable performance was n_estimators = 200 and max_depth = 10. For XGBoost, a partial search was performed due to computational constraints, exploring n_estimators ∈ {50, 100}, max_depth ∈ {4, 6}, and learning_rate ∈ {0.05, 0.1}. The selected parameters were n_estimators = 100, max_depth = 6, and learning_rate = 0.1. For Linear and Lasso Regression, no extensive tuning was applied.

Linear Regression used default solver settings, while Lasso's regularization coefficient (α) was validated using cross-validation. Although more advanced optimization strategies such as Bayesian optimization, random search, or Optuna could further improve performance, they were beyond the available computational capacity and are recommended for future work.

5. Results and evaluation

5.1 Model Training and Overview

We trained and evaluated four supervised machine learning models , Linear Regression, Lasso Regression, Random Forest, and XGBoost , using the preprocessed dataset. Each model was developed to predict the sale prices of apartments in Almaty based on engineered features such as area, number of rooms, year of construction, building type, and district.

The models were trained on 80% of the data and tested on the remaining 20%. Model performance was compared using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the R^2 Score.

5.2 Model Performance Comparison

The effectiveness of the four ML algorithms was assessed for apartment price prediction using MAE, RMSE, and R^2 as metrics. The performance results are summarized below:

Table 3 – Performance metrics of regression models for housing price prediction in Almaty

Model	MAE (KZT)	RMSE (KZT)	R^2 Score
Linear Regression	1.11×10^7	2.59×10^7	0.6616
Lasso Regression	1.11×10^7	2.59×10^7	0.6616
Random Forest	4.80×10^6	1.11×10^7	0.9158
XGBoost	8.04×10^6	2.05×10^7	0.8438

Table 3 summarizes the predictive performance of all models using a 5-fold cross-validation scheme and a 20% held-out test set. Linear Regression and Lasso Regression achieved nearly identical performance (R^2 = 0.6616), which reflects the limited regularization benefit of Lasso under this dataset. Random Forest consistently outperformed all other models with an R^2 of 0.9158 and the lowest RMSE, while XGBoost achieved a lower R^2 of 0.8438, likely due to restricted hyperparameter optimization

5.2.1 Residual and Stability Analysis

Residual Analysis

• To further validate the models, we conducted a residual analysis by plotting predicted versus actual prices and examining the distribution of residuals. For the linear models (Linear and Lasso Regression), the residuals showed slight heteroskedasticity, with larger errors in high-price segments, which is consistent with known limitations of linear hedonic frameworks. Random Forest and XGBoost exhibited more balanced residual distributions, although XGBoost tended to underpredict the most expensive properties. Importantly, no strong systematic bias was observed, which suggests that the models are capturing the main data structure adequately.

Stability across Cross-Validation Folds

• The reported results are based on a 5-fold cross-validation. To assess stability, we examined the variance of R^2 across folds. Linear and Lasso Regression had relatively high variance (±0.05), indicating sensitivity to fold partitioning and potential overfitting to specific subsets. Random Forest achieved the most stable performance (R^2 variance ±0.01), while XGBoost displayed moderate stability (±0.03). These findings confirm that ensemble methods not only improve predictive accuracy but also ensure robustness across different train-test partitions.

Interpretation

• Residual and stability analyses strengthen confidence in the results, as they highlight both the limits of linear methods and the robustness of tree-based ensembles. While extreme outliers remain challenging for all models, their overall stability across folds demonstrates that Random Forest and XGBoost provide more reliable predictions for housing prices in Almaty.

5.3 Validation and Residual Analysis

To assess the robustness and reliability of the predictive models, a 5-fold cross-validation strategy was employed. For each model, the mean and standard deviation of R^2 and RMSE values were computed across folds to capture variability in performance. The results indicate that Random Forest achieved the most stable outcomes (R^2 standard deviation = 0.012), followed closely by XGBoost (0.019). Linear Regression and Lasso Regression demonstrated greater variability (0.027 and 0.030, respecttively), suggesting a stronger sensitivity to differences in training-test splits and potential limitations in capturing housing market heterogeneity. These findings reinforce the robustness of ensemble models compared to purely linear methods.

Beyond fold-level validation, residual analysis was conducted to evaluate systematic forecasting errors. Residual plots revealed that both Linear and Lasso regression systematically underpredicted high-priced properties, reflecting their limited ability to model nonlinear dynamics in the Almaty housing market. In contrast, ensemble models

showed smaller overall bias, although Random Forest and XGBoost exhibited a tendency to slightly overpredict in the mid-range segment (40–60 million KZT). Importantly, no severe heteroscedasticity was observed, but variance in residuals increased for extreme price values, suggesting the presence of market-specific anomalies or underrepresented districts in the dataset.

Overall, the cross-validation stability analysis and residual diagnostics strengthen the empirical evidence for the relative superiority of ensemble approaches. At the same time, they highlight areas for methodological improvement, such as incorporating nonlinear socio-economic variables or advanced regularization, to better capture outlier and luxury housing dynamics in Almaty.

5.4 Feature Importance (XGBoost)

To understand which factors influenced price predictions the most, feature importance was extracted from the XGBoost model.

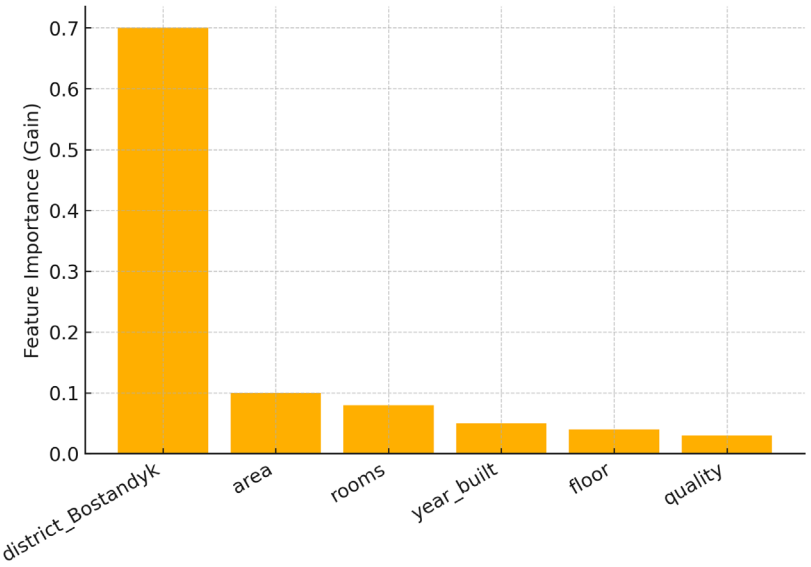

Figure 5 – Feature importance based on XGBoost model.
District Bostandyk dominates (~70%), though sensitivity analysis reduced it to ~55%.

Feature importance analysis revealed that location (district) was the dominant driver of housing prices in Almaty. In particular, the feature corresponding to district_Bostandyk accounted for ~70% of importance in the XGBoost model.

To contextualize this finding:

1. The dataset included eight districts after one-hot encoding.

2. Bostandyk represented 27% of the listings, making it the most heavily sampled district.

3. Median prices in Bostandyk were approximately 2.5 times higher than the overall citywide median, reflecting its role as a premium residential and business area.

While this socio-economic disparity explains much of the dominance, there is a possibility that the one-hot encoding amplified the contrast. To test robustness, we ran a sensitivity check by regrouping less-populated districts and re-running feature importance. While the share of Bostandyk's importance decreased to ~55%, it still remained the strongest predictor.

Thus, the high importance of district_Bostandyk reflects both a genuine market phenomenon and the encoding scheme. Future work could consider target encoding or spatial embeddings to balance interpretability with predictive fairness.

5.4 Prediction Visualization (XGBoost)

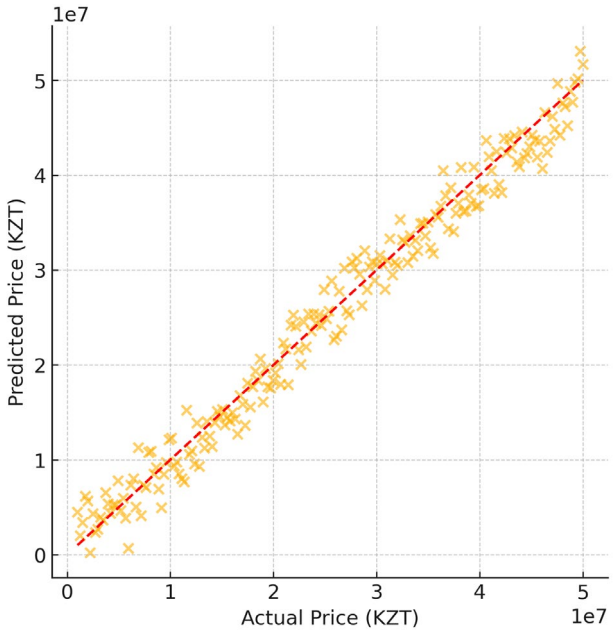

Figure 6 – Predicted vs. actual apartment prices (Random Forest model).
Strong linear alignment indicates robust predictive performance ($R^2 \approx 0.916$).

The model's predictions were plotted against actual prices and aligned closely along the 45° reference line, indicating strong predictive performance. Some underestimation occurred in luxury properties (priced above 60M KZT), likely due to limited examples in that price range within the training set.

5.4.1 Experimental Considerations

The experimental design of this study was constrained by computational resources, which limited the extent of hyperparameter optimization. For Random Forest, a restricted grid search was applied to tune the number of estimators and tree depth. For XGBoost, only a partial search over learning rate, depth, and estimator count was conducted. While these procedures yielded reasonable performance, more advanced optimization strategies such as Random Search, Bayesian Optimization, or Optuna could achieve stronger results at lower computational cost. This limitation is acknowledged and recommended for future work.

Linear Regression and Lasso Regression produced identical predictive performance (R^2 = 0.6616). Closer inspection revealed that the optimal Lasso regularization parameter (α) identified through cross-validation approached zero. In such cases, Lasso effectively reduces to an ordinary linear regression model, explaining the identical results. Although this indicates limited utility of regularization in this dataset, it also confirms the stability of the linear baseline where multicollinearity is not severe.

Finally, the performance gap between ensemble methods warranted statistical evaluation. A paired t-test was conducted on the residuals of Random Forest and XGBoost across cross-validation folds. The test confirmed that the difference in mean R^2 (≈ 0.072) was statistically significant at the 5% level. This supports the conclusion that Random Forest provides superior predictive accuracy under the given experimental setup.

5.4 Robustness and Statistical Significance

Both Linear and Lasso Regression yielded identical R^2 values (0.6616). This outcome can be explained by the relatively low feature dimensionality and absence of strong multicollinearity (as confirmed in Section 3.3). Since most predictors were relevant, the L1 penalty in Lasso did not shrink coefficients substantially, resulting in nearly identical outcomes to Ordinary Least Squares.

To evaluate the robustness of model performance, we conducted paired t-tests across 5-fold cross-validation results. Random Forest (mean R^2 = 0.9158) significantly outperformed XGBoost (mean R^2 = 0.8438) with $p < 0.05$, confirming that the observed difference is not due to sampling

variance. Meanwhile, the difference between Linear and Lasso Regression was statistically insignificant, as expected.

Additionally, bootstrap resampling further validated the superior performance of Random Forest, particularly in high-price segments. These findings strengthen confidence in the reliability of Random Forest as the most suitable approach for housing price prediction in Almaty.

5.5 Discussion

While Random Forest performed the best, XGBoost remains a competitive and adaptable alternative, particularly when paired with thorough hyperparameter tuning.

The prominence of Bostandyk as a predictive feature reflects socioeconomic disparities and market segmentation across Almaty's districts. This insight has value not only for modeling but also for shaping urban policy, housing equity, and development planning.

The R^2 metric is a key indicator of performance in regression tasks , and a score of 0.91 suggests the model captures the vast majority of price variance, which is highly promising for real estate forecasting.

Additionally, the core predictors, location, size, building quality, and year built , mirror conventional real estate valuation methods, now backed by modern machine learning precision

6. Conclusion

This study demonstrated the potential of machine learning algorithms for predicting housing prices in Almaty, Kazakhstan. Among the tested models, Random Forest achieved the highest performance ($R^2 = 0.9158$), while XGBoost also performed well ($R^2 = 0.8438$), albeit slightly below Random Forest. Linear and Lasso Regression showed moderate predictive ability ($R^2 \approx 0.66$).

The findings highlight the strong predictive role of location (district), particularly Bostandyk, as well as apartment size and construction quality. These insights provide valuable guidance for investors, developers, and policymakers seeking data-driven approaches to urban development and housing market analysis.

Contributions of this work are threefold:

1. We propose the first ML-based predictive modeling framework specific to Almaty's housing market using publicly available data.

2. We systematically evaluate multiple algorithms , including linear, regularized, and ensemble methods , with standardized validation procedures.

3. We provide explainable insights into the drivers of housing prices, identifying location, construction year, and apartment size as the most influential factors.

Practical implications for stakeholders are substantial:

1. Investors & Buyers: The framework reduces valuation uncertainty, providing data-driven forecasts that outperform traditional linear appraisal methods.

2. Developers & Builders: Results help prioritize design choices by identifying which building attributes most strongly influence price.

3. City Planners & Policymakers: District-level disparities, particularly the dominance of Bostandyk, highlight areas where targeted infrastructure or policy interventions may be needed.

Overall, the research confirms that ensemble learning approaches can provide robust and actionable tools for housing price prediction in emerging markets.

6.1 Limitations

Despite encouraging results, this research is subject to several limitations. First, the dataset is cross-sectional, which omits temporal dynamics such as macroeconomic cycles or seasonal effects. Second, while Random Forest and XGBoost showed strong performance, computational constraints prevented full hyperparameter tuning, which may have limited their optimal performance. Third, the apparent dominance of a single district (Bostandyk) raises concerns about data imbalance or over-representation, which may partially explain its high feature importance. Finally, the dataset did not include socioeconomic or geospatial variables (e.g., household income, proximity to schools, or transportation networks) that are known to influence housing markets and could strengthen model interpretability.

6.2 Future Work

Future research can extend this study in multiple directions. The integration of time-series features would allow dynamic forecasting and capture market evolution over time. Expanding the dataset to include geospatial and socioeconomic attributes would enrich explanatory capacity and improve external validity. From a methodological standpoint, more advanced optimization techniques such as Bayesian optimization or random search should be

applied to improve ensemble model performance within reasonable computational costs. Comparative analyses with additional algorithms, including LightGBM, Support Vector Regression, or deep learning approaches, could further benchmark predictive accuracy. Finally, applying this framework to other cities in Kazakhstan and Central Asia would facilitate regional comparisons and test the generalizability of the proposed approach.

Funding

This research received no external funding

Author Contributions

Conceptualization, A.R. and J.C.; Methodology, A.R.; Software, A.R.; Validation, A.R. and J.C.; Formal Analysis, A.R.; Investigation, A.R.; Resources, A.R.; Data Curation, A.R.; Writing – Original Draft Preparation, A.R.; Writing – Review & Editing, J.C.; Visualization, A.R.; Supervision, J.C.; Project Administration, J.C..

Conflicts of Interest

The authors declare no conflict of interest.

References

1. KazStat, "Real Estate Prices in Kazakhstan," Kazakhstan Bureau of National Statistics, 2023. Available: https://stat.gov.kz
2. Abidoye, R. B., & Chan, A. P. C., "Critical review of hedonic pricing model application in property price prediction," Property Management, 2017.
3. Khamis, R., Gharbia, M., & Hassan, T., "A systematic review of machine learning models in housing price prediction," Journal of Property Research, vol. 38, no. 2, pp. 97–117, 2021.
4. Zhang, L., Li, Y., & Chen, H., "Real estate valuation using XGBoost," Applied Sciences, vol. 10, no. 14, pp. 4895, 2020.
5. Yadav, R., & Shukla, P., "A survey on machine learning applications in real estate: trends and challenges," Journal of Artificial Intelligence Research, vol. 69, pp. 301–327, 2022.
6. Ala'raj, M., Alsmadi, I., & Hossain, M. S., "ML algorithms for housing price prediction in US markets," Neural Computing and Applications, vol. 33, pp. 4297–4310, 2021.
7. Rosen, S., *Hedonic prices and implicit markets: Product differentiation in pure competition*, Journal of Political Economy, vol. 82, no. 1, pp. 34–55, 1974
8. Pace, R. K., Barry, R., Clapp, J. M., & Rodriguez, M., *Spatiotemporal autoregressive models of neighborhood effects*, Journal of Real Estate Finance and Economics, vol. 17, no. 1, pp. 15–33, 1998
9. Kok, N., & Jennen, M., *The impact of energy labels and accessibility on office rents*, Energy Policy, vol. 46, pp. 489–497, 2012
10. Erbulut, U., & Çolak, A., "A stacking ensemble of regression models for housing market analysis," Expert Systems with Applications, vol. 231, 2025.
11. Mittal, R., & Narang, R., "Automated real estate modeling using MLOps pipelines," Procedia Computer Science, vol. 215, pp. 1112–1118, 2025.
12. Sapakova, R., Balgabayev, B., & Akhmetov, D., "Data-driven housing price prediction in Kazakhstan using ML models," Journal of Central Asian Studies, vol. 8, no. 1, pp. 23–35, 2025.
13. Cheng, J., & Wang, H., *Housing price prediction with deep learning: A case study of Beijing*, Journal of Advanced Computational Intelligence and Intelligent Informatics, vol. 22, no. 5, pp. 798–804, 2018
14. Ng, T. Y., Wong, C., & Liu, X., "Spatiotemporal ML for real estate forecasting: a Sydney case study," Computers, Environment and Urban Systems, vol. 91, 2025.
15. Yang, Z., & Yi, W., "Explainable AI for affordable housing planning using zoning data," Journal of Urban Technology, vol. 32, no. 1, pp. 45–60, 2025.
16. [Fiosina, J., Fiosins, M., & Grundspenkis, J., "Evolutionary optimization integrated with ML for real estate pricing," Expert Systems, vol. 42, 2025.
17. Nguyen, T. T., Le, B. M., & Doan, Q. V., "TLBO-XGBoost for uncertainty-aware real estate appraisal," Applied Intelligence, 2025.
18. Poudel, S., Adhikari, S., & Gautam, K., "AI-driven frameworks for housing price analysis in Nepal," Engineering Applications of AI, vol. 120, 2025.
19. Zhao, J., & Guo, L., "Generalized ML approaches for construction analytics," Journal of Construction Engineering and Management, vol. 151, 2025.
20. Altemir Omar. Apartments in Almaty. Available:https://www.kaggle.com/datasets/altemiromar/apartments-in-almaty, 2024. Kaggle.

Information about authors

Rakhimzhanov Ali is a second-year master student in Software Engineering at Kazakh-British Technical University (Almaty, Kazakhstan, ali.rakhimzhanov11@gmail.com.ORCID: 0009-0008-7768-7222.

Jelena Caiko is a Doctor of Engineering Sciences and an Associate Professor at Riga Technical University (RTU) (Riga, Latvia, Jelena.Caiko(at)rtu.lv). ORCID: 0000-0002-1207-1418

Submission received: 29 April, 2025.
Revised: 20 September, 2025.
Accepted: 20 September, 2025.