➔ *Regular Research Paper – NS*

Predicting Housing Prices in Istanbul Using Explainable Artificial Intelligence Techniques

Hale UYSAL
Burdur Mehmet Akif Ersoy University, Institute of Social Sciences, Deparment of Management Information System, haleyldrn@gmail.com, ORCID: 0000-0003-2997-2236

Adnan KALKAN
Burdur Mehmet Akif Ersoy University, Bucak Zeliha Tolunay School of Applied Technology And Business, Department of Management Information Systems, adnankalkan@mehmetakif.edu.tr, ORCID: 0000-0002-2270-4100

Abstract

Accurate prediction of housing prices in dynamic markets like Istanbul is crucial for stakeholders in the real estate industry, yet traditional models often lack transparency and interpretability. This study addresses this gap by integrating artificial intelligence (AI) with explainable artificial intelligence (XAI) techniques to predict housing prices in the Istanbul housing market. Utilizing a comprehensive dataset containing 25,154 entries and 37 features from the Kaggle platform, we employed several machine learning models, including Random Forest Regressor, Linear Regression, KNeighbors Regressor, Decision Tree Regressor, Gradient Boosting Regressor, and Ridge Regressor. Rigorous data preprocessing steps—such as handling missing values, outlier detection, and encoding categorical variables—were meticulously performed to ensure data quality. The Random Forest model, optimized through hyperparameter tuning, achieved the highest performance with an R^2 score of 0.8683 on the test set. To enhance model interpretability, XAI methods like SHAP and LIME were utilized, revealing that gross square meters and location (specifically, districts like Kadıköy and Sarıyer) significantly impact housing prices. These findings align with existing literature and offer actionable insights for policymakers and industry professionals. This research underscores the importance of combining AI with XAI to develop transparent, reliable models, thereby advancing data-driven decision-making in the real estate sector.

Keywords: *Housing Price Prediction, Machine Learning, Random Forest, Ridge Regression, Explainable AI, Real Estate Market.*

1. INTRODUCTION

Artificial intelligence and machine learning technologies have fundamentally reshaped industries that depend on data-driven decision-making processes. The real estate sector, particularly in dynamic urban markets like Istanbul, is no exception. The application of AI to predict housing prices offers significant benefits by leveraging vast amounts of data to provide accurate, timely insights. These insights are crucial for investors, policymakers, and real estate developers, enabling them to navigate the complexities of housing markets more effectively. However, a critical challenge that remains is the opaque nature of many AI systems, often referred to as "black box" models. These models, while effective, lack transparency in their decision-making processes, making it difficult for stakeholders to fully trust and interpret the

results. This lack of interpretability has hindered the widespread adoption of AI in high-stakes areas like real estate valuation, where understanding the rationale behind predictions is as important as the predictions themselves.

This study seeks to address these challenges by integrating XAI techniques into the AI modeling process for housing price prediction in Istanbul. XAI provides transparency by allowing users to see how various features influence the output of the model. This transparency fosters greater trust in the model, which is crucial for real estate stakeholders who rely on these tools for making informed decisions. By incorporating XAI, this study aims to enhance the decision-making process, making it more transparent and understandable while maintaining the accuracy provided by advanced AI techniques.

The housing market is inherently complex, influenced by a myriad of economic, social, and environmental factors. In a city like Istanbul, which is marked by rapid urbanization, fluctuating property values, and varying socio-economic conditions across districts, the challenge of predicting housing prices becomes even more intricate. Traditional statistical methods often struggle to capture the non-linear relationships present in the data, leading to inaccurate or overly simplistic predictions. Machine learning models, particularly those utilizing ensemble methods like Random Forest, have shown great promise in addressing these challenges by modeling complex, non-linear interactions between variables (Breiman, 2001). However, despite their predictive power, the "black box" nature of these models creates uncertainty for users. As such, there is a growing need for AI systems that are not only accurate but also interpretable.

A key objective of this study is to enhance the transparency of AI models used for housing price prediction through the application of XAI techniques such as SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations). These techniques have proven effective in other domains by breaking down the decision-making process of complex models into understandable components. For instance, SHAP values provide insights into how much each feature contributes to the prediction, offering a clear view of the model's inner workings (Lundberg and Lee, 2017). Similarly, LIME generates local approximations of complex models, making it easier to explain individual predictions (Ribeiro et al., 2016). By employing these tools, this research aims to bridge the gap between accuracy and interpretability, offering a comprehensive approach to housing price prediction.

The significance of this study lies not only in improving the accuracy of housing price predictions but also in fostering greater trust in AI systems. This is particularly important in real estate, where decisions often involve significant financial stakes and long-term investments. By providing explainable predictions, this research aims to offer actionable insights for industry professionals and policymakers. For example, by understanding which factors most strongly influence housing prices, developers can tailor their projects to meet market demand, and policymakers can create more informed housing policies.

In summary, the contributions of this research extend beyond predictive modeling to include the enhancement of model transparency and trustworthiness. This is achieved by integrating cutting-edge XAI techniques into the AI framework, thus making the models not only more accurate but also more interpretable. The following sections will delve into the literature on AI and XAI applications in housing price prediction, the methodology used in this study, the results and their implications, and recommendations for future research.

2. RELATED WORKS

Traditional approaches to assessing and predicting housing prices are dependent on data to a great extent. In relation to this, Çılgın et al. (2023) explain that the accuracy of a machine

learning model is commensurable with the amount of data collected, processed and understood. Similarly, Jafary et al. (2022) demonstrate that the synthetic nature of the datasets to build these models enhances the efficiency of Automated Valuation Models by leaps and bounds. For instance, Morali and Yılmaz (2020) mentioned spatial market segregation for housing in Istanbul as an example of localized price prediction and confirmed that localized prices influence predictions. These types of studies in one way or the other underscore the need to pay attention to accurate and large datasets when building the machine learning models for prediction.

To be able to classically analyze and forecast housing prices, there is a need to investigate how different variables behave and how they can be combined. Fuzzy regression models have been suggested by Sarip et al. (2016) and support of Sarip and Hafez (2015), who emphasize on fuzzy tradingsystems working well with such synthesis of data as housing prices. alpır and Özkan (2018) present the potential and advantages of ANFIS in the processes of real estate valuation. These methods overcome the restrictions of the models which incorporated such uncertainties and ambiguities that naturally arise during real estate transactions into the models thereby enhancing prediction capability.

Machine learning techniques such as Random Forests and Artificial Neural Networks are quite common in prediction of housing prices. Traditional multiple regression modelling is effectively improved upon by the Random Forest technique when working with complex datasets according to Čeh et al. (2018). As Lee and Park (2020) do, it is possible to also incorporate visual property features in models thanks to the use of the deep learning approaches by which great steps are taken than in the common ways where such elements are omitted.

Hedonic pricing models are commonly applied in housing price determinations. Arslanlı in his study, proposed a hedonic model for Istanbul containing many variables of housing and the surrounding neighborhood (Arslanlı, 2020). Keskin and Watkins (2016) contend that the analysis and delineation of spatial housing submarkets can enhance the predictive power of statistical models. Ayan and Eken (2021) highlight the necessity of local studies in predicting prices, as they investigate price bubbles prevalent in certain locations in Istanbul.

There is ample literature regarding how AI may enhance existing processes for improving real estate valuations. Abidoye and Chan (2017) study valuers' attitudes towards artificial intelligence proposing that AI is capable of addressing issues facing valuers and traditional methods of valuation. Odunfa et al. (2021) support the implementation of AI analytical solutions when conducting market analysis and forecasting trends within real estate valuation practice. However, the opaque nature of AI models, dubbed 'black box', generates issues of trust and transparency.

The explainability of AI is crucial for building trust in predictive models. Lundberg and Lee (2017) introduce SHAP values to enhance model transparency by quantifying each feature's contribution to the prediction. Ribeiro et al. (2016) develop LIME to explain model predictions at a local level. Despite their effectiveness, the application of XAI techniques in housing price prediction remains limited. There is a need to address the transparency and interpretability of AI models to increase their adoption in the real estate sector.

Despite the fact that several studies have focused on various modeling approaches for data quality improvement, still very few authors have used XAI techniques in the housing price prediction field in order to make the model more interpretable. This paper attempts to address this shortcoming by embedding XAI techniques in the prediction modeling process and thereby increasing the accuracy of predictions of housing prices in Istanbul while ensuring the explanation of the prediction is possible.

3. METHOD

3.1. Dataset

In this study, the dataset used provides information on a detailed database regarding the Istanbul housing market. The source of the dataset was Kaggle and it consisted of 25154 rows and 37 columns. In each row is a description of a certain property located on certain areas in Istanbul where as the columns contain various property attributes. Among the major variables is included 'price', 'GrossSquareMeters', 'district', 'BuildingAge', 'NumberOfRooms', 'UsingStatus' among others. The scope and the diversity of the data provide an opportunity to undertake an in-depth analysis of the property market in a number of areas in Istanbul.

Because the dataset covers all the areas within Istanbul counties, it makes it a useful tool to study the differences in housing prices by region. It's very important for the research as it analyzes how different price movements according to districts and how these along with the demographic and economic indicators helped achieved the overall results of the study.

3.2. Data Preprocessing

Data preprocessing is a process whereby a number of activities are performed to improve the quality of the data in a dataset at the time of the model development cycle. In the course of conducting this research, some aspects of the concern in the treatment of issues having regard to the missing data, outliers as well as the categorical variables affecting the dataset were given particular importance.

3.2.1. Handling Missing Data

Some variables in the dataset were found to have missing values. More specifically, these values were in the ItemStatus column which concerned some missing data values. To treat this problem, the mean imputation method was used to replace the missing data. Thus filling missing values these moderately explains and accelerates the performance of the model and the accuracy as well.

3.2.2. Detection and Removal of Outliers

Identifiable as a process undertaken in this study, outliers are defined as data points in the dataset that do not adhere to the pattern established and thus would have a negative effect on the model's' accuracy. For example in this study, outliers were especially observed on the GrossSquareMeters and price_per_sqmt columns. Price per square meter values lower than 3,500TL and greater than 35,000TL were inferred as outliers and hence purged from the data set. Such cleaning of data is highly critical in enhancing the uniformity in the model.

3.2.3. Encoding Categorical Variables

For instance, district and ItemStatus are categorical variables that were available in the dataset, so they were encoded in a way that could be acceptable to the machine learning software. Through the one-hot encoding technique, categories were converted to numerical data by primary and secondary flags which enhanced the efficiency of the model.

3.2.4. Data Preparation Process

After the data processing phase, the dataset was ready for operation and analysis. The correction of missing values and outliers, appropriate encoding of categorical variables, and

organization of the dataset raised the quality of the dataset and prepared it for the next stage of model creation.

3.3. Model Development Process

Machine learning models applied during this study are aimed at predicting the cost for houses in the real-estate market of Istanbul. In the modeling process, some activities such as data partitioning, choice of models, hyperparameters tuning and evaluation of models are done.

3.3.1. Splitting Training and Test Data

The dataset was spilt into training which had 80% while the rest 20% would be used for testing. The training set was exclusively used for the model to learn while the evaluation of the overall model performance was done in the test set. This separation is as such very crucial as it determines how well the model can be generalized to other data. In addition, a 5-fold cross validation method was used to assess the model trained in each partition. This technique is useful in having an understanding of how the model fairs when a particular data segment is used.

3.3.2. Model Selection

Selection of models was for their appropriateness in regression analysis and handling of complex relational patterns. Random Forest was selected because of its robustness due to the ensemble nature of active learning and less likely to overfit. Linear Regression was used a baseline model due to its clarity and ease of use. Other models included KNeighbors Regressor, Decision Tree Regressor, Gradient Boosting Regressor, and Ridge Regressor, each offering unique advantages for comparison.

- ***Random Forest Regressor:*** Random forest is a machine learning technique in which several decision trees predict the target variable and normalised outputs of all decision trees predict the target variable. It increases the accuracy of the model by utilizing the output obtained from several trees each of whom arrived at a different decision about the outcome of the prediction. This approach has been selected, because it helps in improving the accuracy and decreasing the possibility of overfitting. Because each of the trees uses different combinations of features, the performance of the model can remain high even when presented with highly complicated datasets. In particular with high dimensional data, the Random Forest nicely models complicated interactions in the data and shows high performance across different new datasets (Tokmak, 2023; Biau et al., 2008).

- ***Linear Regression***: Is the most simplistic as well as a very basic type of regression model mostly used and seeks to find a straight line connecting various independent variables and one dependent variable. Using the linear equation of $Y=aX+b$, the coefficient of each variable is attached to a certain independent variable in order to determine the extent of effect of the influence variable. This approach is popular due to its parsimonious nature and ease of explanation as it works effectively where there exists linear models in the data. Despite this splendor, it is inadequate to model such data that is non-linear (Ng et al., 2018; Asha, 2022; Schneider et al., 2010).

- ***KNeighbors Regressor:*** K-Nearest Neighbours (KNN) relies on the data around the nearest neighbours to make future predictions. Data points are examined in the feature space of individual data points with the help of the closest data points and the predictions are made based on the average calculation. KNN is a basic method that can be very efficient when data is sparse. Unfortunately, when there is a lot of data, it

seems to be overwhelming in terms of cost and low in quality of the results achieved (Ng et al., 2018; Asha, 2022; Okolo, 2010).

- ***Decision Tree Regressor***: The decision tree model makes certain decisions by splitting the data into branches. At each node, the data is split based on a certain feature, and predictions are formed at the leaves. The flexible structure of the model allows it to work with both categorical and continuous variables. Decision trees are highly advantageous in terms of interpretability and can yield successful results, especially in small datasets. However, as the trees deepen, the tendency toward overfitting can negatively affect the model's overall performance (Saravanakumar et al., 2013; Amro et al., 2021; Ursenbach et al., 2019).

- ***Gradient Boosting Regressor***: Gradient Boosting is an algorithm based on sequentially training weak learners and minimizing errors. The model tries to correct these errors in the next step by learning from the errors. This iterative learning process contributes to the model providing high accuracy. Gradient Boosting offers particularly strong performance in complex and large datasets (Biau et al., 2019; Mienye & Sun, 2022; (Yamamoto et al., 2022; Bentéjac et al., 2020). However, computational cost is high, and training can take a long time (He et al., 2017; Hosen & Amin, 2021; Rathnayake et al., 2023).

- ***Ridge Regressor***: Ridge Regression aims to prevent overfitting by adding a regularization term to the classical linear regression model. While minimizing the difference between the predicted values and the actual values, the model also shrinks the coefficients. Thus, if there is multicollinearity among the independent variables, Ridge Regression addresses this situation and increases the model's generalization ability. This model is preferred when there are many independent variables in the data and strong relationships exist among these variables (Zou, 2020; Vatcheva vd., 2016; Xin ve Khalid, 2018; Herawati vd., 2022).

The rationale behind the selection of these algorithms is that each algorithm has specific advantages over other algorithms for certain data structures and conditions. During the study, comparisons were made between the models for accuracy, generalization, and interpretability, and the one that performed the best was used.

3.3.3. Hyperparameter Optimization

Hyperparameter tuning was performed for all of the models. This step is an essential process that is focused on increasing the model's accuracy. In particular, the values of n_estimators, max_depth, min_samples_split, max_features were fine-tuned for the Random Forest model. This step was performed in order for the model to not be subject to overfitting. The list of the best hyperparameters for the Random Forest Model is given in Table 1.

Table 1. Hyperparameter values of Random Forest Model

Hyperparameter	Value
n_estimators (Number of trees)	1000
min_samples_split (Minimum samples to split)	2
max_samples (Maximum samples)	4000
max_features (Maximum features)	5
max_depth (Maximum depth)	50

3.3.4. Model Performance Metrics

To evaluate the performance of the models, metrics such as R^2 Score, Mean Absolute Error (MAE), and Mean Squared Error (MSE) were used. The development of each model's accuracy was evaluated based on these perspectives :

- ***R^2 Score:*** This performance metric indicates how well the model explains the dependent variable. It takes values between 0 and 1; the closer it is to 1, the better the model's performance (Chicco et al., 2021).

$$R^2 = 1 - \frac{\sum(y_{actual} - y_{predicted})}{\sum(y_{actual} - y_{mean})} \quad (1)$$

- ***Mean Absolute Error (MAE):*** It represents the average of the absolute differences between the predicted values and the actual values. The smaller the error, the higher the model's prediction accuracy ((Chai & Draxler, 2014; Hamid et al., 2023).

$$MAE = \frac{1}{n}\sum | y_{actual} - y_{predicted} | \quad (2)$$

- ***Mean Squared Error (MSE):*** It expresses the mean of the squared differences between the predicted values and the actual values. Since MSE gives more weight to larger errors, it penalizes the magnitudes of the errors (Chai & Draxler, 2014; Prasad et al., 2022).

$$MSE = \frac{1}{n}\sum(y_{actual} - y_{predicted})^2 \quad (3)$$

3.4. Application of Explainable Artificial Intelligence (XAI) Techniques

While the development of approaches and models in machine learning is rapidly advancing, understanding internal structures of these models and their interpretations has increasingly become necessary, especially within areas with high consequences decision-making. Such domain is the housing market, which is complex and dynamic, where the prediction results churned out by the model will be more credible if the model is interpretable. In this context, we wanted to understand how such models can be revisited in our case using Explainable Artificial Intelligence methods (Uysal, 2023; Uysal & Köse, 2024). For this reason, XAI approaches based on LIME and SHAP were applied. The purpose of selecting these methods is that to both sides provide explanations of the model that are not dependent on training data and interpret the model on global and local scale. SHAP is known to be introduced as a game-theoretic approach which aims to decompose the overall model interpretability by verifying the contribution of each input feature in the prediction (Lundberg & Lee, 2017). However LIME is a method which proposes to use simple and interpretable models constructed around the predicted outcome to determine the variables responsible for a given prediction (Ribeiro, Singh, & Guestrin, 2016).

The SHAP and LIME methods as described in the literature are a great support in terms of interpretability of any model and have been properly applied in different domains. For instance, SHAP proposed by Lundberg and Lee (2017) helps to elucidate the importance of every feature in the prediction context through a straightforward, uniform, and comprehensive way. In the same way, LIME which was described by Ribeiro et al. (2016) is referred to as a very effective and flexible tool for illustrating the reasons for the focus being on the local prediction. These studies demonstrate how effective SHAP and LIME are in enhancing the understandability of complex machine learning models.

Research by Lundberg and Lee (2017) established that SHAP values elicit transparency in the prediction made but do not provide any improvements to mass predictive models. In this context, the use of SHAP and LIME techniques minimizing the uncertainties characteristic to the so-called 'black box' increases the level of people's confidence in the model and its predictions. For that reason, the combination of SHAP and LIME within this work was chosen with the aim to positively influence the results but also improve the clarity of the analyzed phenomena.

4. RESULTS AND DISCUSSION

This section will present the results obtained from the study in regard to other studies within the literature where necessary and consider the overall conclusions of the study. Furthermore, the sections that concern the scope of the study as well as the section of the recommendations will appear.

4.1. Evaluation of the Findings

The reason why the Random Forest model was superior to the other methods is that it is capable of detecting complex relationships among variables. This is due to combining different decision trees in one model, which decreases the variance in predictions and increases the efficacy of the model (Breiman, 2001). As has been seen before, the aligning variable 'GrossSquareMeters' comes top of the hierarchy. This is as one anticipates because the size of a property will have a relatively high influence on how much it costs. The impacts of the districts Medium on the other hand such as Kadikoy and Sarıyer suggest the importance of a location as Causal effects on which Kauko (2003) placed emphasis in regard to spatial differentiation of housing price. Table 2 provides a comparison of the performance of different models:

Table 2. Comparison of Model Performance

Model	R^2 Score	MAE	MSE
Ridge	0.7859	656,074.22	1,028,708 TL
Linear Regression	0.7853	657,005.55	1,030,132 TL
KNeighbors Regressor	0.7211	673,484.75	1,174,031 TL
Decision Tree Regressor	0.7792	548,965.17	1,044,593 TL
Gradient Boosting Regressor	0.8164	614,284.65	952,578 TL
Random Forest Regressor	0.8508	515,041.29	858,661 TL

As these findings suggest this is why hyperparameter optimization is very effective with regards to the Random Forest model as it shows the extent of the positive effect after optimization. On the basis of the developed model, on the test dataset, the random forest model R^2 score was calculated to be 0.868. The value of MAE however was 502,272 TL and the MSE value 650,682,600 TL. From these results, the conclusion is that the model is capable of working with high precision. Stating the factors affecting prediction of housing prices, gross square meters drove munimum value as the best weighing variable on random forest model. The R^2 was calculated in the range of 95% CLA: [0.85, 0.89] ($P<0.005$) with a straight forward effect of hyperparameter optimization on the model performance analysis. This is an advancement in Shay's argument since a more sophisticated model means the model is able to work on complex data structures and non-linear relationships. Making hyperparameter optimization for the random forest model's parameters also minimized overfitting risks which enhanced the relative efficiency of the model. This result correlates with previous research in the literature and points

towards the need for hyperparameter tuning (Hastie et al., 2009; Bergstra & Bengio, 2012). Table 3 shows the performance of the Random Forest model on the training and test sets:

Table 3. Performance values of the Random Forest Model for training and test set

Metric	**Value**
R^2 Skoru (Training)	0.943
R^2 Skoru (Test)	0.854
MAE (Training)	280,984.26 TL
MAE (Test)	512,926.18 TL
MSE (Training)	213,715,643,252 TL
MSE (Test)	723,642,229,551 TL

Furthermore, it was observed that model performance improved with hyperparameter optimization. Table 4 details the performance of the Random Forest model after optimization:

Table 4. Performance values of the Random Forest Model with Hyperparameter Optimization

Metric	**Value**
R^2 Skoru (Training)	0.9603
R^2 Skoru (Test)	0.8683
MAE (Training)	343,435 TL
MAE (Test)	502,272 TL
MSE (Training)	343,546,286
MSE (Test)	650,682,600

These results further establish that the Random Forest model was more efficient after performing hyperparameter tuning. The R-squared value is 0.868 with the test data set and Mean Absolute Error was 502,272 TL, which means the model did a fair job in predictive modeling. The MSE was determined at 650,682,600 TL. The Random forest model found gross square meters to be the variable that had the most effect on the prediction of the housing prices. So results are further the same as what was presented on the study particularly by Çılgın et al 2023, which claimed that when data is improved prediction of housing prices also becomes enhanced. Further, Jafary et al. 2022, in one of their studies on predicting price of houses utilizing Automated Valuation Models pointed out that data quality played a great role in the success rate of the price predictions.

4.2. Contributions of XAI Techniques

The XAI techniques used in the research, such as SHAP and LIME, have made the model's decision-making processes more transparent and understandable. SHAP values have illuminated the overall functioning of the model by showing which features the predictions rely upon. Figure 1 presents a graphical summary of the SHAP values calculated for each attribute by the model. It was found that the variable with the most significant impact is GrossSquareMeters. While this variable emerged as the most important factor in price predictions, it was also observed that

regions like Kadıköy and Sarıyer have a meaningful effect on prices. Particularly with the assistance of SHAP values, it becomes clearly evident on which data points the model exerts more influence.

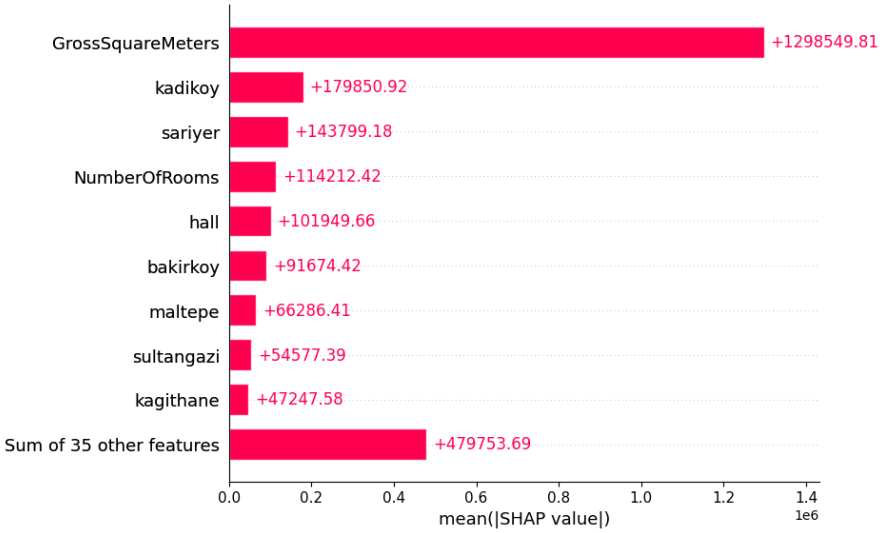

Figure 1. Bar graph showing SHAP values

LIME analyses have also facilitated the explanation of model predictions at the local level. For instance, the factors underlying the predicted price for a specific property were clarified using LIME, making it clear which features the prediction was based upon. Figure 2 presents the local prediction explanation generated by LIME for a property located in the Kadıköy district. According to this analysis, the property's gross square meter area contributes the most significant positive effect to the prediction. Given that the square meter value is specified as 435, possessing such a large area has a substantial impact on the increase in price. Additionally, the property's location in the Kadıköy district also positively affects the price. Since Kadıköy is a valuable and prestigious area in Istanbul, the property's presence in this district favorably influences its valuation.

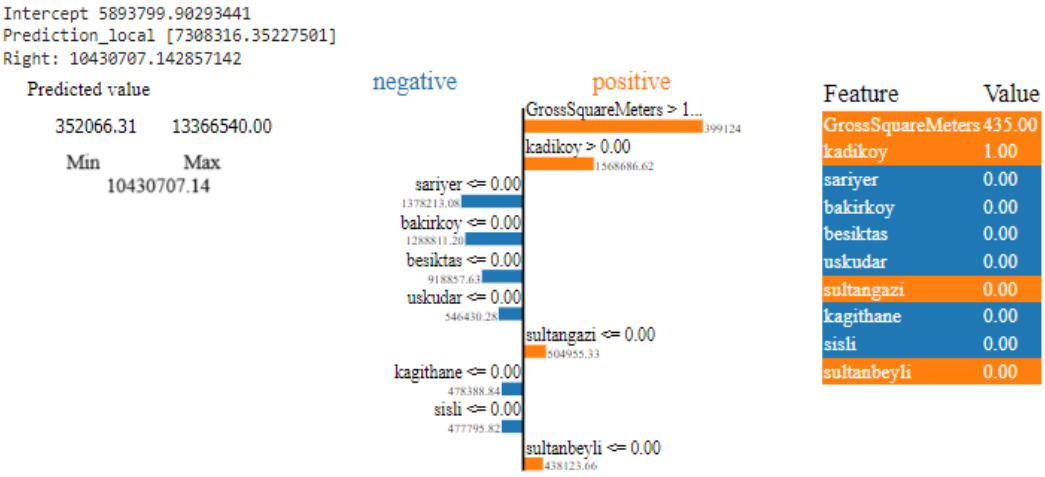

Figure 2. LIME Analysis Result

The ability of XAI techniques not only to enable the model to make accurate predictions but also to explain on which data these predictions are based has increased the model's reliability. These techniques have contributed to decision-makers' increased trust in AI-based systems, allowing them to make more informed decisions.

4.3. Comparison with Literature

When the obtained findings are compared with other studies in the literature, they generally show consistency. The Random Forest model developed by Breiman (2001) has provided high accuracy in housing price prediction by effectively capturing nonlinear relationships. The success of this model stems from its ability to manage complex interactions and nonlinear relationships among variables in the dataset. In our study, the accuracy provided by the Random Forest model aligns with the advantages noted by Breiman (2001), demonstrating that the model successfully handles the complexity in the real estate market.
LIME, defined by Ribeiro, Singh, and Guestrin (2016), and SHAP, developed by Lundberg and Lee (2017), stand out as effective tools for enhancing model explainability. In our study, the integration of these XAI techniques enabled the models' decision processes to become more transparent and understandable. This supports the emphasis by Ribeiro et al. (2016) and Lundberg & Lee (2017) on how increasing model transparency strengthens user confidence and model adoption.

According to Molnar (2022), the explainability of machine learning models is critical for decision making and builds users' confidence in such models. By the integration of SHAP and LIME techniques in the current study, the model's degree of interpretation increased thus allowing more robust decisions by the real estate practitioners and investors as elaborated by Molnar.

Bourassa, et al. (2010) mention that the complexity and quality of information influences Real estate valuation and therefore should not be ignored. Concerning this investigation which concerns on data quality and data scope affecting the accuracy of models, our conclusions are consistent with those of Bourassa et al (2010). This was similarly seen in our study whereby inclusion of broad and good quality datasets boosted the model performance.

Kauko (2003) examined how spatial housing market prices fluctuate and stressed including regional factors as one of the elements in making forecasts. The results of this research revealed that continuity of housing market processes is regionally dependent. Our finding on the interrelation between the housing prices explained by SHAP that was affected by regions like Kadikoy and Sarıyer supports the finding by Kauko (2003). Still, in our case, such an effect was treated in more detail, and regional characteristics in the model were more clearly formulated due to the increased overall translucency of the model with XAI methods.

To sum up, the present study has achieved findings similar to those of related studies in the literature and has managed to add to the existing body of knowledge on the development of more efficient and implementable AI frameworks within the real estate domain by focusing on the model's explainability and transparency through XAI methods.

4.4. Limitations of the Study

There are several limitations of this study. First of all, the dataset utilized in the research restricted to housing information only available in Istanbul, making it difficult to transfer findings to other cities or countries. Also, there was a problem with some variables within the dataset since there was missing data which was done mean imputation to fill. This situation may negatively effect the model's overall performance.

Another limitation is that hyperparameter optimization used to enhance model performance can be time-consuming and costly, especially with large datasets. Future studies are recommended to employ different techniques to make this optimization process more efficient.

5. CONCLUSIONS AND RECOMMENDATIONS

5.1. General Conclusions of the Research

This study encompasses a model development process aimed at predicting housing prices using artificial intelligence and explainable artificial intelligence techniques to improve data-driven decision-making processes in the Istanbul housing market. Various machine learning algorithms were employed, and their performances were evaluated in terms of accuracy and transparency. The analyses showed that the Random Forest model outperformed the other models. The model's R^2 score was calculated as 0.9603 on the training set and 0.8683 on the test set, indicating high overall performance. The high accuracy rate of the Random Forest model can help real estate professionals and investors make more informed decisions, reducing uncertainties in the market.

Another significant contribution of the research is that XAI techniques enable more transparent decision-making processes by increasing the explainability of machine learning models. Through SHAP and LIME analyses, it was clearly explained which data the model's predictions were based on. Consequently, variables such as gross square meters, district, and the number of rooms emerged as the most important factors determining housing prices. Additionally, it was understood that the impact of square meter size on price is the most pronounced, and districts like Kadıköy and Sarıyer also make significant contributions.

These findings are consistent with other studies in the literature and provide important contributions to understanding the dynamics of the housing market in Turkey. Considering the advantages that AI models offer to decision-makers in the sector, it was concluded that incorporating XAI techniques allows for more transparent and reliable decision-making processes.

5.2. Recommendations for Policymakers and Industry Professionals

This study enables professionals in the real estate sector to make more accurate price predictions, allowing them to develop more resilient strategies against market fluctuations. Specifically:

- ***Gross Square Meters:*** It emerges as the most important variable in determining housing prices. It is recommended to place greater emphasis on gross square meter size by setting minimum square meter requirements in new housing projects and integrating them into planning processes.

- ***Regional Differences:*** Significant price differences have been observed among different regions of Istanbul. The higher housing prices in central and prestigious areas (e.g., Kadıköy and Sarıyer) highlight the impact of investments in social and infrastructure amenities on prices. This situation can offer strategic opportunities for investors and real estate developers.

- ***Importance of XAI:*** In addition to AI-based predictions, the explainability of these predictions is also important. Transparent and understandable models instill confidence not only in decision-makers but also in customers and investors. In this context, it is recommended to integrate XAI techniques when using AI models in the

industry. Thus, price predictions will become more reliable, and the adoption of AI applications in the sector will increase.

5.3. Suggestions for Future Studies

Although this research has obtained significant results in the Istanbul housing market, there are areas for improvement in future studies:

- ***Expanding Datasets:*** The dataset used in the study contains housing data only from Istanbul. Future studies can perform more extensive analyses using datasets from different cities or countries, increasing the generalizability of the results.
- ***Integration of Additional Data Sources:*** Integrating additional data sources such as satellite data, environmental factors (air quality, green spaces, etc.), or demographic information can enhance the predictive power of models. The integration of such data sources may allow for a more comprehensive prediction of housing prices.
- ***Advanced XAI Techniques:*** Using more advanced XAI techniques in the future can increase the explainability and accuracy of models. Such techniques can contribute to the wider adoption of AI systems by ensuring that decision-makers have more confidence in model predictions.
- ***Hyperparameter Optimization:*** Although the hyperparameter optimization process used in the research improved the model's performance, it can be time-consuming with large datasets. Future studies are recommended to use more efficient optimization techniques.
- ***Handling Missing Data:*** Filling missing data in datasets is a critical step that can affect model performance. Using advanced data imputation techniques can reduce the impact of missing data, allowing for more reliable results.

REFERENCES

Abidoye, R. and Chan, A. (2017). Valuers' receptiveness to the application of artificial intelligence in property valuation. Pacific Rim Property Research Journal, 23(2), 175-193. https://doi.org/10.1080/14445921.2017.1299453

Amro, A., Al-Akhras, M., Hindi, K., Habib, M., & Shawar, B. (2021). Instance reduction for avoiding overfitting in decision trees. Journal of Intelligent Systems, 30(1), 438-459. https://doi.org/10.1515/jisys-2020-0061

Arslanlı, K. (2020). Analysis of house prices: a hedonic model proposal for istanbul metropolitan area. Journal of Design for Resilience in Architecture and Planning, 1(1), 57-68. https://doi.org/10.47818/drarch.2020.v1i1004

Asha, G. (2022). Linear regression analysis theory and computation. Quing International Journal of Innovative Research in Science and Engineering, 1(2), 39-57. https://doi.org/10.54368/qijirse.1.2.0002

Ayan, E. and Eken, S. (2021). Detection of price bubbles in istanbul housing market using lstm autoencoders: a district-based approach. Soft Computing, 25(12), 7957-7973. https://doi.org/10.1007/s00500-021-05677-6

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3), 1937-1967. https://doi.org/10.1007/s10462-020-09896-5

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of machine learning research, 13(2).

Biau, G., Devroye, L., & Lugosi, G. (2008). Consistency of random forests and other averaging classifiers. Journal of Machine Learning Research, 9(9).

Bourassa, S. C., Cantoni, E., & Hoesli, M. (2010). Predicting house prices with spatial dependence: A comparison of alternative methods. Journal of Real Estate Research, 32(2), 139-160.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32

Chai, T. and Draxler, R. (2014). Root mean square error (rmse) or mean absolute error (mae)? – arguments against avoiding rmse in the literature. Geoscientific Model Development, 7(3), 1247-1250. https://doi.org/10.5194/gmd-7-1247-2014

Chicco, D., Warrens, M., & Jurman, G. (2021). The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. Peerj Computer Science, 7, e623. https://doi.org/10.7717/peerj-cs.623

Čeh, M., Kilibarda, M., Lisec, A., Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. Isprs International Journal of Geo-Information, 7(5), 168. https://doi.org/10.3390/ijgi7050168

Çilgin, C., Gökşen, Y., Gökçen, H. (2023). The effect of outlier detection methods in real estate valuation with machine learning. İzmir Sosyal Bilimler Dergisi, 5(1), 9-20. https://doi.org/10.47899/ijss.1270433

Jafary, P., Shojaei, D., Rajabifard, A., Ngo, T. (2022). A framework to integrate bim with artificial intelligence and machine learning-based property valuation methods. Isprs Annals of the Photogrammetry Remote Sensing and Spatial Information Sciences, X-4/W2-2022, 129-136. https://doi.org/10.5194/isprs-annals-x-4-w2-2022-129-2022

Hamid, A., Nawi, W., Lola, M., Mustafa, W., Malik, S., Zakaria, S., … & Abdullah, M. (2023). Improvement of time forecasting models using machine learning for future pandemic applications based on covid-19 data 2020–2022. Diagnostics, 13(6), 1121. https://doi.org/10.3390/diagnostics13061121

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: Springer.

He, T., Heidemeyer, M., Ban, F., Cherkasov, A., & Ester, M. (2017). Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. Journal of Cheminformatics, 9(1). https://doi.org/10.1186/s13321-017-0209-z

Hosen, S. and Amin, R. (2021). Significant of gradient boosting algorithm in data management system. Engineering International, 9(2), 85-100. https://doi.org/10.18034/ei.v9i2.559

Kauko, T. (2003). Residential property value and locational externalities: On the complementarity and substitutability of approaches. Journal of Property Investment & Finance, 21(3), 250-270.

Lee, C. and Park, K. (2020). Using photographs and metadata to estimate house prices in south korea. Data Technologies and Applications, 55(2), 280-292. https://doi.org/10.1108/dta-05-2020-0111

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

Mienye, I. and Sun, Y. (2022). A survey of ensemble learning: concepts, algorithms, applications, and prospects. Ieee Access, 10, 99129-99149. https://doi.org/10.1109/access.2022.3207287

Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Leanpub.

Morali, O. and Yilmaz, N. (2020). Spatial heterogeneity in istanbul housing market: a geographically weighed approach. Pressacademia, 7(4), 298-307. https://doi.org/10.17261/pressacademia.2020.1304

Ng, S., Chew, Y., Ch'ng, P., & Ng, K. (2018). An insight of linear regression analysis. Scientific Research Journal, 15(2), 1. https://doi.org/10.24191/srj.v15i2.9347

Odunfa, V., Fateye, T., Adewusi, A. (2021). Application of artificial intelligence (ai) approach to african real estate market analysis opportunities and challenges. Advances in Multidisciplinary & Scientific Research Journal Publication, 29, 121-132. https://doi.org/10.22624/aims/abmic2021p9

Okolo, A. (2010). Transformation of independent variables in polynomial regression models. Global Journal of Mathematical Sciences, 8(1). https://doi.org/10.4314/gjmas.v8i1.50810

Prasad, P., Dubey, V., & Sharma, A. (2022). Surface roughness prediction of aisi 304 steel in nanofluid assisted turning using machine learning technique. Key Engineering Materials, 933, 13-24. https://doi.org/10.4028/p-wwb643

Rathnayake, N., Rathnayake, U., Dang, T., & Hoshino, Y. (2023). Water level prediction using soft computing techniques: a case study in the malwathu oya, sri lanka. Plos One, 18(4), e0282847. https://doi.org/10.1371/journal.pone.0282847

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.

Sarip, A. and Hafez, M. (2015). Fuzzy logic application for house price prediction. International Journal of Property Sciences, 5(1), 1-7. https://doi.org/10.22452/ijps.vol5no1.3

Sarip, A., Hafez, M., Daud, M. (2016). Application of fuzzy regression model for real estate price prediction. Malaysian Journal of Computer Science, 29(1), 15-27. https://doi.org/10.22452/mjcs.vol29no1.2

Saravanakumar, D., Ananthi, N., & Devi, M. (2013). An approach to automation selection of decision tree based on training data set. International Journal of Computer Applications, 64(21), 1-4. https://doi.org/10.5120/10755-5500

Seagraves, P. (2023). Real estate insights: is the ai revolution a real estate boon or bane?. Journal of Property Investment & Finance, 42(2), 190-199. https://doi.org/10.1108/jpif-05-2023-0045

Schneider, A., Hommel, G., & Blettner, M. (2010). Linear regression analysis. Deutsches Ärzteblatt International. https://doi.org/10.3238/arztebl.2010.0776

Tokmak, M. (2023). Determination of Maternal Health Status Risk by Machine Learning Methods1. Academic Analysis and Discussions in Engineering, 75.

Ursenbach, J., O'Connell, M., Neiser, J., Tierney, M., Morgan, D., Kosteniuk, J., … & Spiteri, R. (2019). Scoring algorithms for a computer-based cognitive screening tool: an illustrative example of overfitting machine learning approaches and the impact on estimates of classification accuracy.. Psychological Assessment, 31(11), 1377-1382. https://doi.org/10.1037/pas0000764

Uysal, I. (2023). Interpretable Diabetes Prediction using XAI in Healthcare Application. Journal of Multidisciplinary Developments, 8(1), 20-38.

Uysal, I., & Kose, U. (2024). Explainability and the Role of Digital Twins in Personalized Medicine and Healthcare Optimization. In Explainable Artificial Intelligence (XAI) in Healthcare (pp. 141-156). CRC Press.

Xin, S. and Khalid, K. (2018). Modelling house price using ridge regression and lasso regression. International Journal of Engineering & Technology, 7(4.30), 498. https://doi.org/10.14419/ijet.v7i4.30.22378

Vatcheva, K., Lee, M., McCormick, J., & Rahbar, M. (2016). Multicollinearity in regression analyses conducted in epidemiologic studies. Epidemiology Open Access, 06(02). https://doi.org/10.4172/2161-1165.1000227

Yamamoto, F., Ozawa, S., & Wang, L. (2022). Efl-boost: efficient federated learning for gradient boosting decision trees. Ieee Access, 10, 43954-43963. https://doi.org/10.1109/access.2022.3169502

Yalpır, Ş. and Özkan, G. (2018). Knowledge-based fis and anfis models development and comparison for residential real estate valuation. International Journal of Strategic Property Management, 22(2), 110-118. https://doi.org/10.3846/ijspm.2018.442

Zou, H. (2020). Comment: ridge regression—still inspiring after 50 years. Technometrics, 62(4), 456-458. https://doi.org/10.1080/00401706.2020.1801257