Computational Economics (2023) 62:1215–1232
https://doi.org/10.1007/s10614-022-10298-8

A Novel Hybrid House Price Prediction Model

Süreyya Özöğür Akyüz[1] · Birsen Eygi Erdogan[2] · Özlem Yıldız[3] · Pınar Karadayı Ataş[4]

Accepted: 29 June 2022 / Published online: 16 September 2022

Abstract

The real estate sector is evolving and changing rapidly with the increase in housing demand, and new luxury housing projects appear every day. The reliability of housing market investments is largely dependent on accurate pricing.The aim of this study is to introduce a dynamic pricing procedure that estimates house prices using the most important characteristics of a house. For this purpose, a hybrid algorithm using linear regression, clustering analysis, nearest neighbor classification and Support Vector Regression (SVR) method is proposed. Our hybrid algorithm involves using the output of one method as the input of another method for home price prediction to deal with the heteroscedastic nature of the housing data. In other words, the aim of this study is to present a hybrid algorithm that will create different housing clusters from the available data set, classify the houses to which the cluster is unknown, and make price predictions by creating separate prediction models for each class. Housing data collected through manual web scraping of Kadıköy district in Istanbul were used for training and validation of the proposed algorithm. In addition to these data, we validated our algorithm on the KAGGLE house dataset, which covers a wide range of features. The results of the hybrid algorithm were compared using multiple linear regression, Lasso, ridge regression, Support Vector Regression (SVR), AdaBoost, decision tree, random forest and XGBoost regression. Experimental results show that the proposed hybrid model is superior in terms of both Residual Mean Square Error (RMSE), Mean Absolute Value Percent Error (MAPE) and adjusted Rsquare measures for both Kadıköy and KAGGLE housing dataset.

Keywords Housing pricing · Support vector regression · K-means clustering · K-NN classification

Extended author information available on the last page of the article

1 Introduction

Recent studies have shown that there is a rapid increase in housing investments in the construction sector, which plays an important role in the economic development of countries (Ozkan et al. 2012; Cengiz et al. 2019). For this reason, the evaluation of the factors affecting the housing market and the estimation of the price of a house according to its characteristics have started to attract increasing academic interest.

In the last ten years, a rapid growth has been seen in the construction sector in Turkey with large investments in infrastructure and urban renewal projects. Meanwhile, the demand for housing has also increased in Turkey due to the increase in population, changes in lifestyle and rising living standards. Therefore, developing and modernizing the housing pricing mechanism, a task usually undertaken by real estate agents in Turkey today, is of critical importance for all parties involved in the real estate market. Therefore, studies are continuing in Turkey to both predict the sales prices of the houses built and to predict the direction of the general house price index. Our aim in this research is to contribute to these studies by suggesting a hybrid approach that can be used to make price estimation using the properties of the houses.

In Pagourtzi et al. (2003), property prices are determined by two different approaches, referred to as traditional and advanced valuation methods. Traditional valuation methods include the comparable method, investment/income method, profit method, development/residual method, contractor's method, cost method, multiple regression method, and stepwise regression method. Advanced valuation methods, for their part, are used to simulate the decision processes of the market players and predict the change in these decisions. Artificial Neural Networks (ANNs), hedonic pricing method, spatial analysis method, Fuzzy Logic (FL), and autoregressive integrated moving average (ARIMA) methods can be given as examples of advanced valuation methods.

In some recent studies, macroeconomic variables have been used to estimate the housing index, unlike our cross-sectional study, which includes the properties of houses (Aquaro et al. 2021; Bourassa et al. 2019; Milunovich 2020; Case and Shiller 1990). On the other hand, in many studies, the hedonic approach and multiple regression analysis are used, which are also known as hedonic regression (Sasaki and Yamamoto 2018). For a 'one-stop' reference of hedonic approaches, one can look at the review study of Herath and Maier (2010). Some of the studies are mentioned below in order to highlight the similarities and differences with our study.

In Stevenson (2004) earlier hedonic house price model is reassessed by adding the average age of homes in Boston. Using the statistical average age of the houses, his results demonstrated that there is heteroscedasticity in housing data. Stevenson referred to study in Goodman and Thibodeau (1997), who provided evidence that the age of the dwelling is a primary cause of heteroscedasticity. In Bin (2004), a model using Geographic Information System (GIS) data is developed to take heteroscedasticity into account using the location and characteristics of the houses. In this study, a hedonic price function was estimated using semi-parametric regression. The heteroscedastic nature of the housing data inspired us to use a clustering approach

as a preprocessing step of the housing data before we use an appropriate model for price prediction of the houses.

Over the last decade, artificial neural networks decision tree and fuzzy logic methods have also been used to predict house prices. Fan et al. (2006) applied a decision tree approach on the Singapore resale public housing market. The results showed the usefulness of this method in determining the relationship between house prices and housing characteristics. In Selim (2009) the determinants of house prices in Turkey for the whole country are considered, including both urban and rural areas, using both hedonic regression and ANN methods. According to Del Giudice et al. (2017); Siti Norasyikin Abd. Rahman (2019) fuzzy logic system (FLS) is able to handle real estate predictions since it can model relations between independent and dependent variables more accurately. One recent study on housing pricing estimation in Eskişehir, Turkey employed a fuzzy logic approach in which the distances from the house to cultural, educational, medical buildings, transportation systems, in addition to other environmental attributes were taken as the main features of the data (Kusan et al. 2010). In addition to these features, housing sales prices may vary according to the marketing capabilities of real estate agencies. According to Kuşan's view, the hedonic method and multiple regression methods are not enough to estimate house prices and cannot deal with problems such as outliers, non-linearity, discontinuity, and fuzziness. Therefore, in their study, house unit prices were estimated using fuzzy logic where environmental, transportational and regional socio-economic factors were used as independent variables. In our approach, instead of searching for additional variables to explain the heteroscedastic structure in the data set, a hybrid approach in which regression, clustering and classification analysis are combined is proposed.

Some hybrid and smarter approaches in the prediction of housing prices have also arisen over the last decade. For example, in Gerek (2014) the performance of Adaptive neuro-fuzzy (ANFIS) with grid partition and sub clustering models is compared. It was shown that ANFIS with grid partition models performed better than ANFIS with sub clustering models. In Park and Bae (2015) various classification methods are used such as C4.5, RIPPER, Naïve Bayesian, and, AdaBoost to predict if a townhouse would be sold for less or more than the list price using the house characteristics. They compared their classification accuracy performances to other classification methods, demonstrating the RIPPER algorithm's superiority. Recently, ensemble methods have come to the fore as an approach that increases the accuracy of prediction by combining the decisions of different models (Bowen and Buyang 2018; Neloy et al. 2019; Paireekreng and Choensawat 2015). Hybrid algorithms and ensemble learning expressions are sometimes mistakenly used interchangeably. Lu et al. (2017) and Truong et al. (2020) use a so-called "hybrid" fusion of regression models for housing price prediction, but in fact, the estimates from each regression model are averaged to produce the final result. Therefore, the approach used by the researchers is not a hybrid algorithm or a hybrid model, but rather an ensemble learning.

The hybrid algorithm we propose in this study combines both hedonic regression and machine learning techniques, including clustering and classification methods. We prefer to call our approach "hybrid" rather than ensemble because it doesn't integrate

decisions the way ensemble methods do (Bowen and Buyang 2018; Paireekreng and Choensawat 2015; Lu et al. 2017; Truong et al. 2020). Instead, our hybrid approach uses the output of one algorithm as the input of another within the entire schema.

As will be detailed in the 2 section, the contribution of the proposed algorithm can be listed as follows:

- Splits the training data into clusters by incorporating the errors from multiple regression into the k-means algorithm.
- Makes class prediction for test data using K-nn by using the class labels as dependent variable.
- Estimates pricing for each identified cluster using nuSVR.
- Pioneers the use of hybrid algorithms for housing pricing.

In the final step of estimating house prices, we tested our hybrid approach using various regression techniques. These methods include multiple linear regression, Lasso, ridge regression, Support Vector Regression (SVR), AdaBoost regression, decision tree regression, random forest regression (Bühlmann and Yu 2010), and XGBoost regression (Chen and Guestrin 2016). We then compared the performances of these methods with their hybrid versions for each.

Housing pricing data for the Kadıköy/Istanbul region of Turkey were collected through the website of the sahibinden (2019), which serves as a real estate database to the public. In addition, the "The Ames Housing" dataset uploaded to KAGGLE by De Cock (2011) was used to implement our proposed hybrid algorithm. Experimental results show that the proposed approach outperforms existing literature methods for both datasets in terms of prediction.

The remainder of the article is organized as follows: In the next section, we provide an abstract description of the methods used in our hybrid approach. In Sect. 2 we describe our hybrid method, followed by Experiments and Results in Sects. 3 and 3.2 respectively. Finally, we conclude Sect. 4 with a summary and discussion of the advantages of the proposed method.

1.1 Multiple Regression

Linear regression is a well-known method with a wide scope of application areas. The fundamental aim of multiple regression is to learn the relationship between several independent variables and a dependent variable. For instance, a real estate specialist may want to see how the price of a house is affected by the size of the house (in square feet), the number of bedrooms, the average income in the respective neighborhood according to census data, and a subjective rating of the appeal of the house. One might learn that the size of the flat is a better predictor of the price than the floor that it is on.

Multiple linear regression considers the model of the distribution of a continuous type quantitative response variable Y_i of the $i-th$ observation for given explanatory variables $X_{i1}, \ldots, X_{ip}$ as follows (Hastie et al. 2009):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i, \quad (1)$$

where β_0 refers to intercept, $\beta_1, \ldots, \beta_p$ are regression coefficients and ϵ_i denotes the error term which has zero mean and which captures the residual variability. Estimation of these paramaters β_i is done by well-known methods such as least square or maximum likelihood estimation (Vapnik 1998).

In a multicollinearity case, Ridge regression and Lasso are commonly used methods to estimate the coefficients of variables (Hoerl and Kannard 1975). In this study, we compared the prediction performance of the proposed hybrid approach with ridge regression and Lasso because of the multicollinear structure of the data. The most related two independent variables, causing the multicollinearity, are the area of the house and the number of the rooms. Researchers dealing with housing data often observe the effect on the parameter estimation in their attempt to use multiple linear regression for price prediction. Because of the very well-known effect of multicollinearity, for example, the coefficient of the room number may appear negative while the coefficient of the area of the house is positive. Logically, we expect both of them to be positive. That is why we suggest the use of ridge regression and Lasso instead of multiple linear regression. Besides, it is known that even though multicollinearity affects the variances of the coefficients, still may be good at prediction. Therefore, a researcher may also use multiple linear regression instead of ridge regression or Lasso if the primary aim of the study is prediction rather than interpretation of the coefficients.

1.2 K-means Clustering

Clustering is a method that groups similar objects by minimizing the distance within the cluster while maximizing the distance between the clusters. K-means is one of the most preferable methods of clustering techniques, as it minimizes the distance between the centroids of the clusters and the observations within the cluster. During the clustering process, observation/examples are added iteratively until the smallest distance is achieved (Kaufman and Rousseeuw 2009). Here, K refers to the number of clusters given initially to the algorithm as a parameter. In our study, the appropriate number of clusters is determined using the so called Elbow method that uses a grid search algorithm included in scikit learn. The overview of the clustering algorithm can be explained in the following three steps (Hastie et al. 2009):

1. The input space is divided in to K clusters and examples are randomly assigned to the clusters. These serve as initial cluster assignments for the observations.
2. For each data point
   - Find the distance between the example and centroid of the cluster,
   - If the example has the shortest distance to its own cluster (centroid) then leave it else select another cluster.
3. Repeat steps 1 and 2 untill there is no observation left to be moved from one to another cluster.

The common distance measures in K-means algorithm are the Euclidean distance, the Euclidean squared distance and the Manhattan or City distance. In this paper, we used Euclidean distance for K-means method.

1.3 K-NN classifier

K Nearest Neighbor (K-NN) method is the simplest algorithm that can be used both for regression and classification. It is preferred due to its low calculation time and ease of interpretation.

For given training data $X = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \subseteq \mathbb{R}^n \times \mathbb{R}$, and a test point (x_t, y_t), K-NN algorithm predicts the class labels of the test examples by looking at K most similar training examples. In classification cases, it assigns the majority class label (majority voting). Similarly, it assigns the average response for regression. K-NN is also called a non-parametric method as it does not learn an explicit mapping f from the training data but rather uses the training data at the test time to make predictions. It needs an input of K nearest neighbors and the distance function to compute the similarities between the examples. Here we used K as 1. There are several ways to compute distances based on the type of features. For example, Euclidean distance is commonly used for real-valued features, whereas Hamming distance is preferred for binary valued features (Altman 1992). We have used the Euclidian distance.

1.4 Support Vector Regression (SVR)

Support Vector Machines are developed for classification problems in general (Vapnik 1995), and have been extended to regression analysis with a new heading called Support Vector Regression (SVR). The method is linear regression in the sense of hyperplanes determined in SVR, but it is nonlinear in the sense of interpreting the data points in output space using kernel functions. SVR has quickly become popular because it does not have any assumption for the data.

For a given training set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \subseteq \mathbb{R}^n \times \mathbb{R}$, SVR fits the data points into a (hyper-) tube of diameter 2ϵ with $\epsilon \geq 0$. This hypertube can be considered as a regression model which fits the data points with a hyperplane positioned in its center. There are many possible ways to locate a hypertube of diameter 2ϵ. However, there exists an optimal hypertube that contains as many training points as possible. The optimal hypertube can be found by maximizing the distance of observations from the center hyperplane which has the same idea of maximizing margin in SVM principle (Hamel 2011). SVR has two parameters C and ϵ to be optimized where epsilon can be considered as a parameter that affects the accuracy of the solution, but in most cases the solution is required to be as accurate as possible. In order to overcome this problem, ϵ can be included as a part of the optimization problem which turns into a $\nu-$SVR problem where $0 \leq \nu \leq 1$ is the regularization constant in the objective function of the optimization problem below (Hastie et al. 2009):

$$\begin{aligned} \min_{w,b,\xi,\xi^*,\epsilon} \frac{1}{2} w^T w + C(v\epsilon + \frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*)) \\ (w^T \phi(x_i) + b) - y_i \leq \epsilon + \xi_i, \\ y_i - (w^T \phi(x_i) + b) \leq \epsilon + \xi_i^*, \\ \xi_i, \xi_i^* \geq 0, i = 1, \ldots, l, \ \epsilon \geq 0. \end{aligned} \tag{2}$$

Here, w is the normal of the hyperplane, C is the error constant, b is the bias term, l refers to the number of observations in the training set, ξ_i is the $i-th$ slack variable $(i = 1, \ldots, l)$ and $\phi(.)$ is a nonlinear mapping from input space to output space.

In v-SVR, the parameter v is used to determine the proportion of the number of support vectors we desire to keep in our solution with respect to the total number of samples in the dataset. In this study, we used $v-$ SVR and the parameters of $v-$SVR were determined on the training set by using the well-known classical model selection method "k-fold cross validation".

2 Hybrid Algorithm

We develop a novel hybrid algorithm where each decision maker (models) outputs are used as input for the next step to make correction on misclassified observations. In this regard, the proposed algorithm is novel because of the correction steps within the training data points. In other words, models are attached to each other in our hybrid algorithm, where we used multiple regression, k-means clustering, K-NN classifiers and support vector regression. Integration of these methods are explained with an algorithm and a workflow diagram in Fig. 2.

Determining the value of a home is very important to the real estate market. The algorithm we propose can be used to set up a dynamic system that will automatically determine the value of the residences using their properties. In this study, the first step of the hybrid algorithm consists of dividing the data into training and test sets, where regression is applied on the training set and prediction errors are obtained on the test set. It was observed that the application of multiple regression revealed errors in about three categories (error vs. y-estimation), reflecting the heterogeneous nature of the data, illustrated by the simple scatter diagram in Fig. 1. Here, the first category includes the housing prices underestimated by the regression model, the second category includes the closely estimated housing prices, and the third category includes the overestimated housing prices.

These breaks, noted in the scatterplot of errors, inspired the use of a clustering approach on errors to discover training subsets. Indeed, the idea of clustering came from the heterogeneous nature of the available data. It is known that there are some subsets of housing data depending on regional and/or physical factors.

In the second step, the K-means clustering algorithm was applied to *residual error vector* in order to find the number of classes which coincided with the visual results shown by the Fig. 1. The purpose of using the clustering method is to automatically determine the breakpoints of the errors revealed by regression analysis. Training folder labels were updated with new clustering labels using K-means. In the third

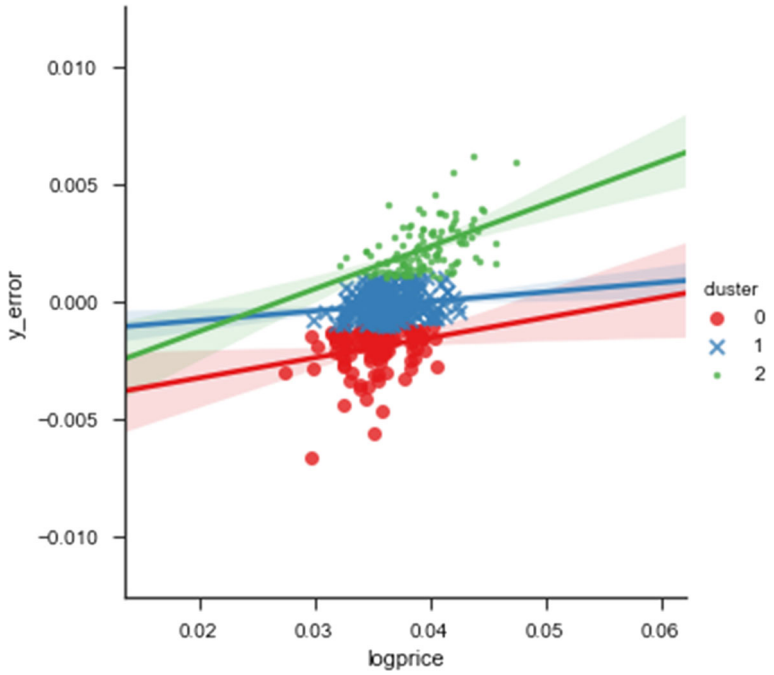

Fig. 1 Error based clustered data visualization

step, K-NN classification was applied to the updated training data to predict the class labels on the test folder. After all class labels were estimated, the training labels updated using the K-means and the test labels predicted using the K-NN model were concatenated. In the final step, a regression method (multiple linear regression, Lasso, ridge regression, Support Vector Regression (SVR), AdaBoost, decision tree, random forest and XGBoost regression) was applied to each cluster separately to predict housing prices. With this proposed hybrid algorithm, it is possible to first classify houses and then predict prices independently for each class. All these steps are presented in Algorithm 1 and Fig. 2.

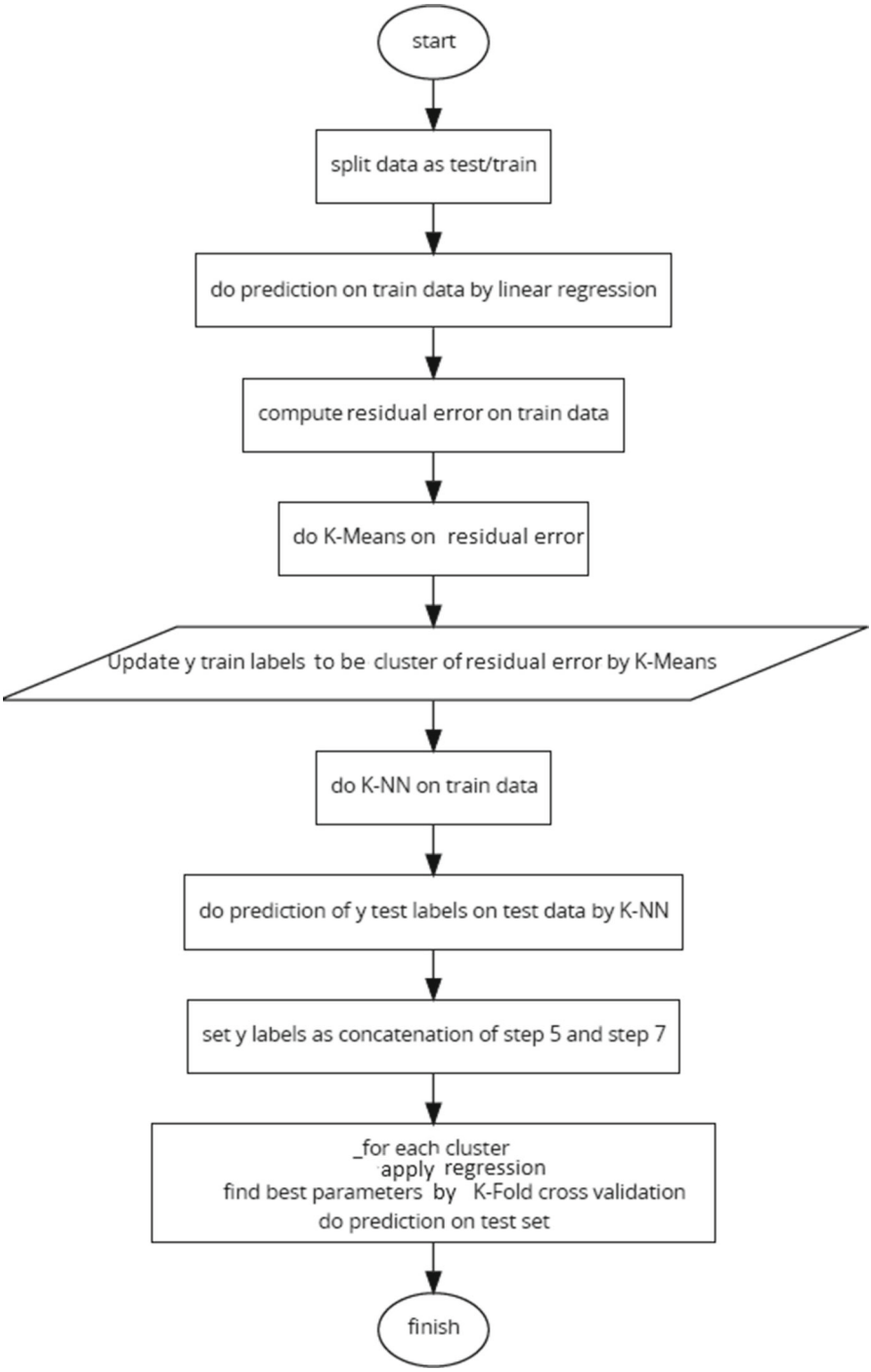

Fig. 2 Flow chart of Algorithm 1

Algorithm 1 Hybrid House Price Prediction Algorithm

Input: Data: (X, Y)
Output: Prediction Performance

```
split the data into Train and Test
do prediction on train data using linear regression
compute residual error vector on train data
do K-Means on residual error vector
set y train labels as the cluster of residual error vector determined using K-Means
do K-NN on train data
do prediction of y test labels on test data using K-NN
update y labels as concatenation of step 5 and step 7
for all cluster do
    apply regression method
    find the best parameters of regression methods (if any) using K-Fold cross validation
end for
do prediction on test set
```

As a summary, first, the class labels were determined for all samples in the data set by integration of K-means and K-NN. Then regression methods like standard linear regression, Lasso, ridge regression, nu-SVR, Support Vector Regression (SVR), AdaBoost regression, decision tree regression, random forest regression, and XGBoost regression were applied to each cluster on a separate run per method for comparison purpose to obtain a final prediction. To find the best parameters of regression methods, k-fold cross-validation was used independently for each cluster. In this step, a holdout validation was also used to test each regression method on the test folder.

The proposed hybrid approach was compared with the standard linear regression, Lasso, ridge regression, nu-SVR, Support Vector Regression (SVR), AdaBoost regression, decision tree regression, random forest regression, and XGBoost regression in terms of Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Rsquare values in Tables 3 and 4. The results are discussed in the next section.

3 Experiments

3.1 Data Sets

In this study, we tested our hybrid approach[1] on two datasets. We collected the first dataset through a commercial platform serving real estate agents and all stakeholders related to the real estate industry in Istanbul, Turkey. Here we have focused on the Kadıköy area. The second dataset (KAGGLE Ames Housing Dataset) is a more universal dataset collected by De Cock (2011) for a data mining competition organized by KAGGLE. More details about the datasets will be given in the following subsections.

[1] https://github.com/pnrkaradayi/HousePricePrediction

3.1.1 Kadıköy Housing Data

We collected 744 observations from 2014-2015 through manual web scrapping from the (sahibinden 2019) website,which serves as a commercial platform for realtors, homeowners and buyers. In our data analysis, typical preprocessing steps were performed to scale the variables, as there are variables measured at different scales. For example, the price of the house, the dependent variable, was very positively skewed corrected by a natural logarithm function. The house characteristics we collect are as follows: square meter of the house (logm2), number of rooms (nroom), number of bathrooms (nbathroom-tr), number of floors of the apartment (floor), total number of floors in the building (nfloor), a dummy variable indicating whether the house is in a closed complex site (insite), and a dummy variable showing the year of sale (year). The properties of the data is given by Table 1.

At the beginning of the study, the data were monitored by bivariate correlations and error terms of the multiple linear regression. It was observed that the parameter estimations were unstable and gave the wrong signs due to the correlations between features. This well-known, so called multicollinearity problem appeared when "house size", "number of the rooms", and "number of bathrooms" were used in the same model. The relationship between the features can be seen in the heat map which reflects the correlations given in Fig. 3.

It is clear in Fig. 3 that the features most correlated with price are house size (logm2), number of bathrooms(nbathroom-tr), and the number of rooms(nroom). In addition, when the correlation values are examined, it is seen that these variables, which are separately related to the price and express the width of the house, are also related to each other. It may be considered unnecessary to include all these variables in the model, but both the size of the house and the way it is partitioned are known as factors affecting the purchasing decision. Therefore, it may be preferable to keep the related variables in the model and use ridge regression or a non-parametric approach to predict house prices.

3.1.2 KAGGLE Housing Data

"The Ames Housing" is a dataset containing 80 different variables (23 nominal, 23 rows, 14 discrete and 20 continuous) associated with housing prices, with approximately 3000 observations sold between 2006 and 2010. These variables, which show the characteristics of the houses in the dataset, were then used to estimate the average price of each house. We selected the first 30 features shown in Table 2 based on their correlation information.

In the data analysis, a typical pre-processing step was performed. The pre-processing procedure applied to the data set are given below:

- Apply encoding and labeling on non-numerical data column.
- Ignore the variables (columns) with more than %90 missing values and for those having low missing values (lower than %30) impute the numerical and categorical variables with median and the most frequent values of the variables respectively.

Table 1 Kadıkoy housing dataset variables

Variable	Type	Desription
m2	Numeric	Square meter of the flat
nroom	Numeric	Number of rooms
nbath-tr	Numeric	Total number of bathroom in the flat
floor	Numeric	Number of floors of the flat
nfloor	Numeric	Total number of floors in the building
insite	Categorical	Whether the house is in a complex or not
year	Categoric	0-1 for two different years of the sale
price	Numeric	Selling price of the house

Fig. 3 Multiple correlation matrix of Kadıköy Dataset

- Create a new variable TotalSqFeet from the surface area of each floor. The surface area of each floor has low correlation with house price; however, when we sum them up, the relationship becomes much stronger.
- The target variable "SalePrice" is right-skewed. Likewise we examined skewness in the rest of numerical variables and used log transformation to fix all of them.
- Calculate the correlation coefficient of variables and consider the top 30 variables based on correlations visualized in Figure 4 where the number of top variables to be considered in the model was decided experimenatly on training folder considering the best training accuracy.

3.2 Results

As described in Sect. 2, data were first clustered using the K-means algorithm and then various techniques such as multiple linear regression, Lasso ridge regression, Support Vector Regression (SVR), AdaBoost regression, decision tree regression, random forest regression, and XGBoost regression were used for each cluster to predict the house prices. Then we compared each method with their hybrid versions in Tables 3 and 4.

The results presented in Tables 3 and 4 are the average of 10 random runs where we get competitive results with the proposed hybrid approach. For each run, we choose the best tuning parameters using cross validation. For example for ν-SVR with Gaussian kernel, the optimum C, ν and γ values are selected from the intervals given by

Table 2 Ames housing dataset variables

Variable	Type	Desription
OverallQual	Categorical	Overall material and finish quality
TotalBsmtSF	Numeric	Total square feet of basement area
CentralAir	Categorical	Central air conditioning
GarageCars	Numeric	Size of garage in car capacity
GrLivArea	Numeric	Above grade (ground) living area square feet
LotArea	Numeric	Lot size in square feet
YearBuilt	Numeric	Original construction date
GarageArea	Numeric	Size of garage in square feet
FullBath	Numeric	Full bathrooms above grade
YearRemodAdd	Numeric	Remodel date (same as construction date if no remodeling or additions)
OverallQual	Categorical	Overall material and finish quality
RoofStyle	Categorical	Roof type
BsmtUnfSF	Numeric	Unfinished square feet of basement area
Neighborhood	Categorical	Physical locations within Ames city limits
MSSubClass	categorical	The building class
GarageType	Categorical	Garage location
1stFlrSF	Numeric	First floor square feet
TotRmsAbvGrd	Numeric	Total rooms above grade (except bathrooms)
GarageFinish	Categorical	Interior finish of the garage
OpenPorchSF	Numeric	Open porch area in square feet
Bedroom	Numeric	Number of bedrooms above basement level
ExterQual	Categorical	Exterior material quality
SaleCondition	Categorical	Condition of sale
KitchenQual	Categorical	Kitchen qualityv
WoodDeckSF	Numeric	Wood deck area in square feet
BsmtFinType1	Categorical	Rating of basement finished area
FireplaceQu	Categorical	Fireplace quality
BsmtQual	Categorical	Height of the basement
Fireplaces	Numerical	Number of fireplaces
ExterQual	Categorical	Exterior material quality
Heating	Categorical	Type of heating
SalePrice	Numeric	The property's sale price in dollars.

$$C = [0, 10, 100, 1000],$$
$$v = [0.1, 0.3, 0.5, 0.7, 1.0],$$
$$\gamma(\text{Gaussian kernel parameter}) = [1e-4, 1e-3, 0.01, 0.1, 0.2, 0.5].$$

It is observed that the proposed hybrid approach achieves better RMSE, MAE, MAPE and adjusted Rsquare values for Kadıköy dataset than the corresponding

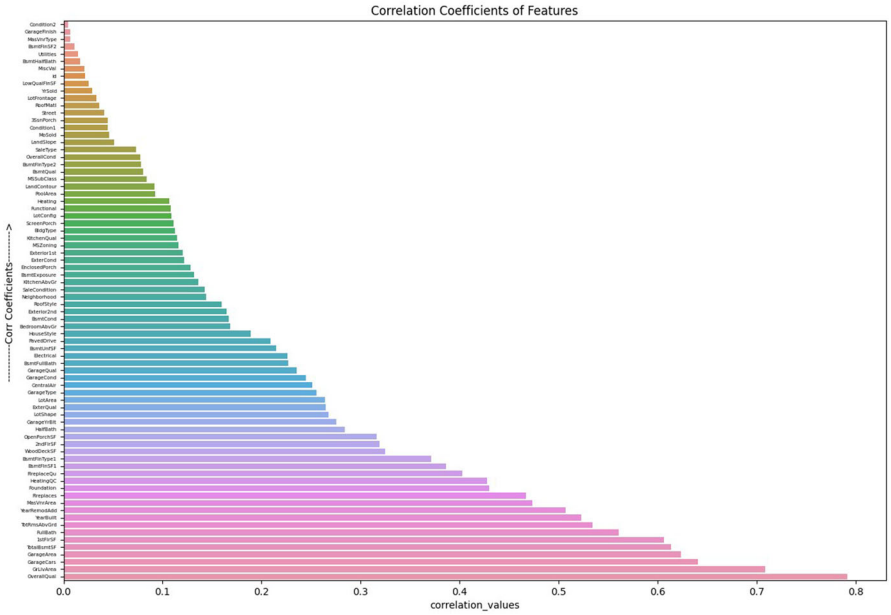

Fig. 4 Correlation coeffficients of variables of KAGGLE dataset

Table 3 The proposed hybrid algorithm versus other standard techniques on Kadıköy dataset

Model	RMSE	MAE	MAPE	Adjusted Rsquare
nu-SVR	0.00168	0.0013	3.421458	0.674176
Hybrid-nu-SVR	0.00127	0.00090	2.401432	**0.779466**
Ridge Regression	0.000007	0.002059	5.561910	0.381333
Hybrid - Ridge Regression	0.000003	0.001332	3.580693	**0.622271**
Lasso	0.000007	0.002059	5.561910	0.445784
Hybrid - Lasso	0.000003	0.001124	2.530773	**0.556012**
Linear Regression	0.000003	0.001681	3.444298	0.654739
Hybrid - Linear Regression	0.000002	0.000910	2.437791	**0.791333**
AdaBoost Regression	0.000004	0.001527	4.158058	0.595008
Hybrid-AdaBoost Regression	0.000002	0.001127	3.035137	**0.701333**
Decision Tree Regression	0.000004	0.001466	3.939253	0.560765
Hybrid - Decision Tree Regression	0.000002	0.001134	3.040145	**0.712667**
Random Forest Regression	0.000003	0.001298	3.488057	0.624213
Hybrid - Random Forest Regression	0.000002	0.000977	2.601959	**0.742000**
XGB Regression	0.000004	0.001436	3.851183	0.584745
Hybrid - XGB Regression	0.000002	0.000934	2.512429	**0.788000**

Table 4 The proposed hybrid algorithm versus other standard techniques on KAGGLE dataset

Model	MSE	MAE	MAPE	Adjusted Rsquare
nu-SVR	0.003210	0.082401	1.874321	0.765231
Hybrid - nu-SVR	0.0025	0.009737	0.831034	**0.874992**
Ridge regression	0.013247	0.100021	1.562312	0.781132
Hybrid - ridge regression	0.074360	0.174530	1.579532	**0.800013**
Lasso	0.010201	0.102253	1.531928	0.821572
Hybrid - Lasso	0.012650	0.134009	1.639761	**0.854535**
Linear regression	0.026258	0.159115	0.892285	0.840945
Hybrid - linear regression	0.024500	0.113578	0.965755	**0.936667**
AdaBoost regression	0.032192	0.137356	1.1467718	0.822798
Hybrid-AdaBoost regression	0.083278	0.203429	1.759382	**0.895833**
Decision tree regression	0.040259	0.141749	1.186507	0.762825
Hybrid - decision tree regression	0.042701	0.147523	1.245084	**0.773333**
Random forest regression	0.021994	0.100294	0.839654	0.864437
Hybrid - random forest regression	0.144709	0.245733	2.115874	**0.896000**
XGB regression	0.023579	0.108746	0.909801	0.854591
Hybrid - XGB regression	0.077155	0.181168	1.560980	**0.925833**

standalone algorithms in Table 3. But for the KAGGLE dataset, the hybrid approach outperformed only nu-SVR and linear regression in terms of RMSE. On the other hand, our hybrid approach results in highly competitive adjusted Rsquare values compared to standalone versions of all methods for both datasets. The low RMSE, MAE and MAPE values for the KAGGLE dataset in the hybrid method may arise because of the imputations made due to missing data and the higher number of attributes in the KAGGLE dataset compared to the Kadıköy dataset.

4 Conclusion

The housing market is crucial for the country's economy and the estimation of housing prices is very important for all stakeholders involved in the purchase and sale of housing. This study aims to estimate house prices with a hybrid algorithm using certain properties of houses. The relationship between house properties and house prices is implicitly modeled through a hybrid learning algorithm rather than a classical statistical learning approach. Our hybrid modeling consists of two steps. In the first step, K-means clustering was used to update the training labels on the residual error vector, derived from linear regression, which is often used for house price prediction. In the second step, the updated training dataset and test dataset were classified and combined using the K-NN algorithm. In the last step, house prices in each cluster were estimated using different methods for comparison purposes. These methods include multiple linear regression, Lasso, ridge regression, Support Vector

Regression (SVR), AdaBoost, decision tree, random forest, and XGBoost regression. Each of these methods is compared with hybrid versions in terms of RMSE, MAE, MAPE, and adjusted Rsquare. Experimental results show that the proposed hybrid approach achieves better results than the standalone versions of each method.

5 Discussion

This hybrid approach can be used to create a dynamic system for home price prediction based on statistical learning. Because the algorithm itself has the ability to update the predictions with the correction step it applies for an additional new observation to the training set. Therefore, it can be configured to make automatic housing price estimation by processing every new information entered into the system, especially through online systems. The location of the houses, nearby social facilities, competition (how many houses are for sale on the supply side, and how many people are looking for a house on the demand side) are thought to be factors that will increase the success of house pricing. However, since the data collected in this study is commercially available on the internet via sahibinden (2019), such additional features could not be reached. To see the effects of these features and to test our proposed hybrid method, we analyzed KAGGLE data with more variables. Experimental results demonstrate the usefulness of our proposed hybrid algorithm as well as the improvement in performance measures as more features are added to the dataset, as shown in Sect. 3.2. As a future study, this hybrid approach could be tested in areas other than home price prediction, such as car price forecasting or cell phone price prediction.

Author Contributions SÖA: Conceptualization, Methodology, Writing- Original draft preparation. BEE: Investigation, Data Collection Methodology, Reviewing and Editing. ÖY: Software, Validation PKA: Feature Engineering and data pre-processing.

Funding The authors have not disclosed any funding.

Declarations

Conflict of interest All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician, 46*(3), 175–185. https://doi.org/10.1080/00031305.1992.10475879

Aquaro, M., Bailey, N., & Pesaran, M. H. (2021). Estimation and inference for spatial models with heterogeneous coefficients: An application to us house prices. *Journal of Applied Econometrics, 36* (1), 18–44.

Bin, O. (2004). A prediction comparison of housing sales prices by parametric versus semi-parametric regressions. *Journal of Housing Economics, 13*(1), 68–84.

Bourassa, S. C., Hoesli, M., & Oikarinen, E. (2019). Measuring house price bubbles. *Real Estate Economics, 47*(2), 534–563.

Bowen Y, Buyang C (2018) Research on ensemble learning-based housing price prediction model. Big Geospatial Data and Data Science 1

Bühlmann, P., & Yu, B. (2010). Boosting. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(1), 69–74.

Case, K. E., & Shiller, R. J. (1990). Forecasting prices and excess returns in the housing market. *Real Estate Economics, 18*(3), 253–273.

Cengiz, S., Atmiş, E., & Görmüş, S. (2019). The impact of economic growth oriented development policies on landscape changes in istanbul province in Turkey. *Land Use Policy, 87*, 104086.

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp 785–794

De Cock D (2011) Ames, iowa: Alternative to the boston housing data as an end of semester regression project. Journal of Statistics Education 19(3)

Del Giudice, V., De Paola, P., & Cantisani, G. (2017). Valuation of real estate investments through fuzzy logic. *Buildings, 7*(1), 26. https://doi.org/10.3390/buildings7010026

Fan, G. Z., Ong, S. E., & Koh, H. C. (2006). Determinants of house price: A decision tree approach. *Urban Studies, 43*(12), 2301–2315.

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

Goodman, A. C., & Thibodeau, T. G. (1997). Dwelling-age-related heteroskedasticity in hedonic house price equations: An extension. *Journal of Housing Research, 8*, 299–317.

Hamel, L. H. (2011). *Knowledge discovery with support vector machines* (Vol. 3). London: Wiley.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction,* (Vol. 2). Springer.

Herath S, Maier G (2010) The hedonic price method in real estate and housing market research. a review of the literature. Tech. rep., WU Vienna University of Economics and Business

Hoerl, A. E., Robert, Kannard BKF., & W,. (1975). Ridge regression: Some simulations. *Communications in Statistics-Theory and Methods, 4*(2), 105–123.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis* (Vol. 344). Wiley.

Kusan, H., Osman, A., & Ozdemir, I. (2010). The use of fuzzy logic in predicting house selling price. *Expert Systems with Applications, 37*(3), 1808–1813.

Lu S, Li Z, Qin Z, Yang X, Goh RSM (2017) A hybrid regression technique for house prices prediction. In: 2017 IEEE international conference on industrial engineering and engineering management (IEEM), IEEE, pp 319–323

Milunovich, G. (2020). Forecasting Australia's real house price index: A comparison of time series and machine learning methods. *Journal of Forecasting, 39*(7), 1098–1118.

Neloy AA, Haque HMS, Ul Islam MM (2019) Ensemble learning based rental apartment price prediction model by categorical features factoring. In: Proceedings of the 2019 11th International Conference on Machine Learning and Computing, Association for Computing Machinery, New York, NY, USA, ICMLC '19, p 350-356, 10.1145/3318299.3318377

Ozkan, F., Ozkan, O., & Gunduz, M. (2012). Causal relationship between construction investment policy and economic growth in Turkey. *Technological Forecasting and Social Change, 79*(2), 362–370.

Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real estate appraisal: A review of valuation methods. *Journal of Property Investment & Finance, 21*(4), 383–401.

Paireekreng W, Choensawat W (2015) An ensemble learning based model for real estate project classification. Procedia Manufacturing 3:3852–3859, https://doi.org/10.1016/j.promfg.2015.07.892, 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015

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*(6), 2928–2934.

Sasaki M, Yamamoto K (2018) Hedonic price function for residential area focusing on the reasons for residential preferences in japanese metropolitan areas. Journal of Risk and Financial Management 11 (3), 10.3390/jrfm11030039

Selim, H. (2009). Determinants of house prices in turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications, 36*(2), 2843–2852.

Siti Norasyikin Abd Rahman MRSI N H A Maimun (2019) The artificial neural network model (ann) for Malaysian housing market analysis. Planning Malaysia Journal 17

Stevenson, S. (2004). New empirical evidence on heteroscedasticity in hedonic housing models. *Journal of Housing Economics, 13*(2), 136–153.

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

Vapnik, V. (1995). *The nature of statistical learning theory.* New York: Springer.

Vapnik, V. (1998). *Statistical learning theory.* Wiley.

wwwsahibindencom (2019) Housing data. https://www.sahibinden.com/

Authors and Affiliations

Süreyya Özöğür Akyüz[1] · Birsen Eygi Erdogan[2] · Özlem Yıldız[3] · Pınar Karadayı Ataş[4]

✉ Süreyya Özöğür Akyüz
sureyya.akyuz@eng.bau.edu.tr

Birsen Eygi Erdogan
birsene@marmara.edu.tr

Özlem Yıldız
yldz.ozlm@gmail.com

Pınar Karadayı Ataş
pinaratas@arel.edu.tr

[1] Faculty of Engineering and Natural Sciences, Department of Mathematics, Bahçeşehir University, Istanbul, Turkey

[2] Faculty of Science, Department of Statistics, Marmara University, Istanbul, Turkey

[3] Big Data Analytics Program, Institute of Science, Bahçeşehir University, Istanbul, Turkey

[4] Faculty of Engineering and Architecture, Department of Computer Engineering, Arel University, Istanbul, Turkey