Computational Economics (2022) 59:549–577
https://doi.org/10.1007/s10614-021-10099-5

Determining the Flat Sales Prices by Flat Characteristics Using Bayesian Network Models

Volkan Sevinç[1]

Accepted: 25 January 2021 / Published online: 6 February 2021

Abstract
There are various factors affecting flat sales prices. Various characteristics of a flat play an important role in determining its sales price. In this study, a machine learning based Bayesian network was built by a restrictive structural learning algorithm using the data collected from 24 randomly selected cities in Turkey. The data consist of the sales prices and various characteristics of a flat such as number of bedrooms, building age, availability of balcony, net area, heating type, mortgageability, number of bathrooms, seller type, presence in a housing estate area and floor location. After the model validity check, a sensitivity analysis was performed for the estimated Bayesian network model and related results were provided. Some of these results indicate that sales prices of flats mostly change depending on the number of bathrooms available. Additionally, number of bedrooms, net area and floor location are also determinative about the sales prices. The lack of significant difference among the sales prices of flats that are sold by construction companies, housing estate agents or property owners is another result obtained.

Keywords Flat sales prices · Real estate sector · Residence sale · Bayesian networks

1 Introduction

The need for residence is an indispensable need for people. Therefore, residence prices have always been a topic of interest in daily life and in various branches of science. In addition, residence prices in a country are one of the factors affecting the economy of the country. Estimating residence prices is very important for different purposes such as local or foreign people who want to buy a house for investment or dwelling in, banks providing mortgage credits, construction companies and real estate agents. Certain properties of a residence play a decisive role in the formation

✉ Volkan Sevinç
vsevinc@mu.edu.tr

[1] Department of Statistics, Faculty of Science, Muğla Sıtkı Koçman University, Kötekli Kampüsü, 48000 Muğla, Turkey

of it sales price. Although there are many factors that affect the sales price of a residence, there is no general agreement on how and to what extent these factors affect the price of the residence. Thus, researchers have been trying to build various models to estimate the sales price of a residence based on its various aspects. Hedonic price models, which are based on the consumer theory suggested by Lancaster (1966), have been being used widely for this purpose for a long time. Hedonic pricing identifies the factors and characteristics affecting the price of an item. Thus, hedonic price models estimate the sales price of a residence considering both its internal characteristics and some external factors that are likely to affect its price. House price estimation with hedonic pricing models is quite popular and repeated in many studies using data belonging to various countries. According to Sirmans et al. (2005), results provided by hedonic pricing models depend mostly on location and time. Moreover, the value that a buyer appraises for various properties of a residence varies considerably according to the social, cultural, economic, climatic conditions of the place where the residence is located. However, hedonic price models assume that buyers have the same amount of desire towards the same attributes of a property at the same level, which is not very possible.

Bayesian networks however have some advantages in determining the effects of the variables. They have also some advantages over the regression models like hedonic pricing. For example, in regression models, regressors, which are used to estimate the dependent variable, are called independent variables, which are expected to be uncorrelated, and possible dependencies among them are ignored. Any change in the value of an independent variable affects only the dependent variable and not any other independent variables in the model. This way of modeling the sales price of a residence based on its various attributes may not be very realistic. Indeed, for example, when the number of bathrooms increases in a residence, it directly causes an increase in the total area of the residence. Bayesian network models, on the other hand, consider the dependency relations among all the variables in the model unless some restrictions are imposed. As the variables exist in a network structure they all can influence each other at various strengths. Another disadvantage of the regression models is that they are not as flexible as Bayesian network models. That is, regression models are additive models taking the sum of the independent variables and equalizing it to the dependent variable. After fitting a regression model, adding some extra independent variables that are found to be significant, may dramatically change the structure and even the accuracy of the model. Bayesian models, however, as they mainly consider the interdependencies among the variables rather than the sum of them, are not that highly affected by new variables being added into the model. Bayesian networks also do not suffer from overparameterization problem very much like regression models do. Another advantage of Bayesian network models is that they can handle missing data more easily than regression models.

In the literature, the number of studies on residence sales prices and containing a Bayesian approach is very few. In a study by Hui et al. (2010) hierarchical Bayesian approach was used to estimate the price of residences in Hong Kong. Giudice et al. (2017) developed a residence price estimation model based on the Bayesian approach. They used Markov Chain Hybrid Monte Carlo method for estimating the

house prices in Naples, Italy. Liu et al. (2018) combined a manually formed causal Bayesian network depending on the expert opinion and a hedonic model to estimate the residence sales price in Nanjing, China.

2 Research Gaps and Contributions of the Study

Flats are the most popular types of residences across the EU. In fact, Eurostat (2018) reports that, 46% of people in the European Union (EU) countries lives in flat type of residences, 18.6% in semi-detached houses and 34.7% in detached houses. According to Turkish Statistical Institute (TurkStat), similar to the EU, flat type residences also are the most popular residence units in Turkey (Turkstat 2011). Therefore, in this study, we aim to estimate the sales prices of flats depending on their various characteristics by using a model other than the classic hedonic pricing models. As we mentioned in the previous section, Bayesian network models have various advantages over regression models. Moreover, there is not a previous study in the literature estimating the sales prices of residences by a Bayesian network model based on a machine learning algorithm. Thus, in this study we develop a Bayesian network model by using a machine learning algorithm that constructs the pattern of the network through structural learning and estimates the sales prices of flats in Turkey based on various characteristics of the flats.

In this paper, definition and properties of Bayesian networks, their visual and probabilistic features, types of Bayesian networks and scoring functions are provided in Sect. 2. Information about some Turkish housing statistics, data collection procedure, the variables used in the study and their types, the levels of the variables and their definitions are presented in Sect. 3. The model selection procedure, the estimated Bayesian network model, the phases of PC algorithm, the validation of the estimated model for both the training data and the testing data and the sensitivity analysis of the model are presented in Sect. 4. Various findings of the estimated Bayesian network and discussions and comparisons about these findings with the results obtained in similar studies are presented in Sect. 5. Finally, conclusions about the study and some suggestions for future studies are provided in Sect. 6.

3 Materials and Methods

3.1 Bayesian Networks

Bayesian networks express the conditional probabilistic relations among variables as a graphical model. Bayesian networks were first proposed by Pearl (1985). Bayesian networks have two main structures as graphical and probabilistic. The graphical structure of Bayesian networks is based on a relationship structure called Directed Acyclic Graph (DAG). A DAG consists of nodes and directed arrows that connect them to each other. Variables in the model are represented by nodes, and arrows indicate the conditional probabilistic relations among the variables. In a DAG, an arrow pointed out from a node can never return to the

same node by following any path of the directed arrows. This property of DAGs arises from the necessity that a causal relation must never return to the node that it started from. In Bayesian networks, the estimation of conditional dependencies is made by the creation of Conditional Probability Tables (CPT) which form the probabilistic part of Bayesian networks.

In Bayesian networks, if two nodes are connected with an arrow, the node from which the arrow is pointed out is called the parent node and the one the arrow is pointed to is called the child node. Nodes that have no parent node are called root nodes, and nodes that have no child node are called leaf nodes. Figure 1 shows an example Bayesian network in DAG structure.

In Fig. 1, for instance, node A is the parent of node B and G, node B is the child of node A and node F. Moreover, it can be concluded that there is no conditional probabilistic relationship between K and E variables since there is no arrow between these nodes connecting them. Node J has no parent node, so it can be called a root node. On the other hand, node E is a leaf node since it has no child node.

In Bayesian networks, there are two approaches to building the DAG structure. The first approach is based on the opinions of experts. In the light of these opinions, the relationships and their directions among the variables are determined. As a result of this approach, there is a causal relationship between a parent node and a child node. Bayesian networks established with this approach are called causal Bayesian networks.

In the second approach, the formation of the network structure is realized by various algorithms applied to the data set. In this way, the structure of the network is created based on machine learning. In Bayesian networks that are based on machine learning algorithms, the directions of arrows refer to probabilistic dependences rather than cause–effect relationship. Such networks are also classified as Bayesian networks based on structural learning.

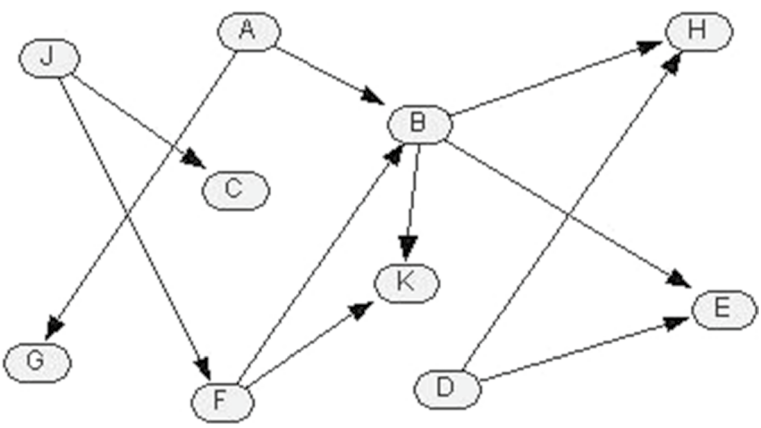

Fig. 1 Bayesian network in DAG structure

3.2 Causal Bayesian Networks

Causal Bayesian networks are created manually based on expert opinion. Causal Bayesian networks allow observing the cause–effect relations and probabilistic dependencies among variables. There is no limitation on the number of variables or how many parent–child relations will be established. These numbers are determined by the decision of an expert in the related area. Still, the Bayesian network structure must have the DAG structure.

3.3 Bayesian Networks Based on Structural Learning

Another approach to estimating Bayesian networks is using a data set. There are many methods for learning Bayesian networks by means of a data set. These methods generally form the structure of the Bayesian network directly from the data set with the help of various machine learning algorithms. Learning methods of Bayesian networks from the data set can be examined under three main topics. These topics are: score-based methods, constraint-based methods and hybrid methods.

Score–based methods first create a variety of network structures that fit the data set and they rank these network structures with respect to some scoring functions such as K2, Bayesian Dirichlet (BD), Log-Likelihood (LL), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Afterwards, the network structure with the highest score is selected as the model. Among these scores, calculation of LL score is as follows (Heckerman et al. 2000).

$$Log - likelihood\left(V_1, \ldots, V_r\right) = -\frac{\sum_{i=1}^{r} log_2 P\left(V_i/model\right)}{rk} \tag{1}$$

where r is the number of cases and k is the number of nodes existing in the model. Calculation of AIC score by (Akaike 1973) can be given as

$$AIC(B|T) = LL(B|T) - |B| \tag{2}$$

Moreover, BIC value by Schwarz (1978) is calculated as follows.

$$\varphi(B|T) = LL(B|T) - f(1/2logN)|B| \tag{3}$$

The fundamentals of constraint-based methods were given by Verma and Pearl (1990). They begin to learn the Bayesian network structure with the whole graphical structure. Afterwards, they apply some conditional probability tests among the nodes and the edges between any nodes that do not have a conditional dependency relation, are eliminated. For more information on Bayesian networks, Jensen and Nielsen (2007), Holmes and Jain (2008) and Spirtes (2010) can be referred.

4 Data Collection

Construction sector is one of the important components of the Turkish economy. According to the population and housing survey by TurkStat (2011), the share of the housing sector in the Gross Domestic Product (GDP) reached 30%. Moreover, 74.9% of the buildings in Turkey have residential qualification. In Turkey approximately 16,514,000 of the residences are flat type of residences while 1,427,867 of them are private workplaces, 304,546 are public workplaces and 546,454 of them are summer houses or seasonally used buildings. There is a total of 19,482,000 households in Turkey. 19,454,000 (99.9%) of these households live in buildings which have residential qualification. Of these residences, 20% consists of single–story buildings, 19.5% consists of two–story buildings, 11.9% consists of three–story buildings and 48.6% consists of four–story or higher buildings (TurkStat 2011). Additionally, according to the results of TurkStat (2011), in Turkey, 1% of the residences have 1 bedroom, 7.3% have 2 bedrooms, 39.9% have 3 bedrooms, 45.1% have 4 bedrooms and 6.6% have 5 or more bedrooms. Moreover, 23.4% of the households in Turkey were built before 1980, 43.5% between 1981 and 2000 and 21.8% in 2001 and later. According to a research by Ajans Press (2018), in the last 10 years a total of 9,361,941 residences have changed hands in Turkey. Furthermore, the number of houses sold in 2008 was 427,111 compared to 1,375,398 in 2018. Therefore, it is possible to say that residence sales have increased by 322% in Turkey during the last 10 years. Flats are the most common residence types in Turkey. The data used in the study were drawn from a property sales website www.sahibinden.com with the help of a code written in R language. The data were collected in October 2019 during various connection sessions to the website. The data belong to 24 randomly selected cities in Turkey. These cities are Adana, Ankara, Antalya, Aydin, Balikesir, Bursa, Denizli, Diyarbakir, Duzce, Erzurum, Eskisehir, Gaziantep, Kayseri, Kocaeli, Konya, Istanbul, Izmir, Mersin, Mugla, Ordu, Sakarya, Samsun, Tekirdag and Tabzon. The data consist of 5000 observations corresponding to 11 variables indicating various features of a flat. The variables were determined based on the results of Lancaster (1966) who suggested consumer theory which led to the development of hedonic price models. In consumer theory, it is stated that utility is related to the attributes of a good. Each house has its own unique set of characteristics. Hence, in hedonic price models, the price factors are identified both by internal characteristic of the good and external factors affecting it. In this study 7 internal characteristics (number of bedrooms, building age, existence of balcony, net area, heating type, number of bathrooms and floor location) and of a flat being sold and 3 external factors (mortgageability, seller type, and flat being within a housing estate) were taken into consideration. All of the variables in the study are categorical types of variables which are either nominal or ordinal. The variables used in the study and their types and definitions are given in Table 1.

Table 1 Nodes used in the study, their levels and definitions

Node		Levels	Definition
1	Number of Bedrooms (ordinal variable)	1	Indicates the number of bedrooms in the flat in addition to 1 living room
		2	
		3	
		4	
		5 and over	
2	Building Age (years) (ordinal variable)	New	Indicates the age of the building which the flat is a part of
		1–4	
		5–10	
		11–15	
		16 and over	
3	Balcony(nominal variable)	Available	Indicates whether or not the flat has a balcony
		Not Available	
4	Net Area (m^2) (ordinal variable)	0–120	Specifies the amount of net area of the flat in meter squares (m^2)
		120–250	
		250 and over	
5	Heating Type (nominal variable)	Combi Boiler	Specifies the heating type used in the flat
		Floor Standing Boiler	
		Air Conditioner	
		Block Heating	
		Coal Stove	
6	Sales Price (TL) (ordinal variable)	0–200,000	Indicates the sales price of the flat in Turkish Lira (TL)
		200,000–400,000	
		400,000–800,000	
		800,000–2,000,0000	
		2,000,000 and over	
7	Mortgageable (nominalvariable)	Yes	Indicates the eligibility of the flat to get mortgage credit from banks

Table 1 (continued)

Node		Levels	Definition
		No	
8	Number of Bathrooms (ordinal variable)	1	Specifies the number of bathrooms in the flat
		2	
		3 and over	
9	Seller (nominal variable)	Real Estate Office	Specifies the type of the seller as a real estate office, a construction company or a property owner
		Construction Company	
		Property Owner	
10	Within a Housing Estate (nominal variable)	Yes	Indicates whether the flat is located in a housing estate area
		No	
11	Floor Location (ordinal variable)	Basement	Specifies the floor on which the flat is located
		Ground Floor	
		1st floor	
		2nd Floor	
		3rd Floor	
		4th Floor	
		5th Floor and Higher	
		Loft	

5 Results

The dataset was randomly divided into two parts in a way that 70% as training data and 30% as testing data. In accordance with the nature of the data used in the study, the structural learning algorithm to be applied was decided to be a restrictive learning algorithm. Thus, no arrow was allowed to be directed from the major target variable "flat sales price" to any other nodes as the flat sales price variable is the dependent variable in the model and it is not supposed to affect any other predictors.

Using the training data, three different structural learning algorithms, which allow restrictions, in GeNIe (2019) software were employed and three different Bayesian network models were estimated based on each of these algorithms named Bayesian Search (BS), Peter-Clark (PC) and Greedy Thick Thinning (GTT). Among these three different Bayesian network models estimated, the selection was made based on the LL scores calculated by GeNIe (2019).

LL score can be defined as a measure of fit of the model to the data or how successful is the model in reproducing the dataset. It is calculated as the logarithm of the probability of the whole dataset, given a specific model structure. This probability usually appears to be a very small value because it is equal to the multiplication of the marginal probabilities of all the cases in the data set. LL score is a value between minus infinity and zero. As LL score decreases, the model becomes less likely to fit the data. Table 2 shows the algorithms experimented during the model estimation stage and their LL scores.

It can be seen in Table 2 that the algorithm having the highest LL score is PC algorithm. Thus, the model estimated by PC algorithm seems to fit to the data better than the ones estimated by the other two algorithms. Hence, the Bayesian network model estimated by PC algorithm was selected to be used in the analysis of the flat sales prices.

5.1 PC Algorithm

PC algorithm is a constraint-based type of algorithm. It was introduced by Spirtes and Glymour (1991). PC algorithm basically performs a series of independence tests among the variables using the data set. Based on the results of these tests, edges between nodes that do not have any dependency relation are eliminated. These tests can be information based such as BIC, ad-hoc or statistical and they decide to remove or not an edge between two nodes based on a p-value they calculate (Tsagris 2019). The phases of PC algorithm are given by Spirtes and Glymour (1991) as follows.

Table 2 Bayesian network algorithms experimented and their Log-Likelihood scores

Algorithm	Log-likelihood score
PC	− 472,037
Greedy thick thinning	− 480,771
Bayesian search	− 481,151.83

Let A_{ab} denote the set of edges adjacent to nodes a or b. Let P_{ab} denote the set of edges that are parents of nodes a or b. Let U_{ab} denote the set of edges on undirected paths between a and b.

(a) On the vertex set V, the complete graph C_{-1} is formed.
(b) For each pair of nodes a, b that are adjacent in C_n:

(i) If $A_{ab} \cap U_{ab}$ does not have cardinality greater than n, go to the next pair of edges adjacent in C_n.
(ii) If A_{ab} has cardinality greater than n, check if a and b are independent conditional on any subsets of $A_{ab} \cap U_{ab}$ of cardinality $n + 1$. If so, delete the edge between a and b from C_n.

Let C_{n+1} be the graph resulting from the procedure above that is applied to each pair of nodes. Continue until a value of $f + 1$ of n is reached such that (ii) is not satisfied for any pair.

(c) For each triple of edges a, b, c such that the pair a, b and the pair b, c are each adjacent in C_f but the pair a, c are not adjacent in C_f, orient $a - b - c$ as $a \rightarrow b \leftarrow c$ if and only if a and c are dependent on every subset of $A_{ac} \cap U_{ac}$ containing b. Construct all graphs consistent with these orientations

For a detailed information about PC algorithm Spirtes and Glymour (1991), Meek (2013) and Tsagris (2019) can be referred.

The estimated Bayesian network model which was visualized in Netica (2019) software is given in Fig. 2. The indicator values behind the bars in the model are percentage values.

5.2 Model Validation

In order to evaluate the model estimation performance for single levels of the dependent variable, ROC curves can be employed. The Area under the Curve (AUC) is the amount of area under the ROC curve which gives the probability of correct estimation. ROC curves are drawn by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity). For the assessment of the overall estimation performances of the models, confusion matrices can be referred. Confusion matrices display the number of correctly and wrongly classified cases by the model. AUC scores for each flat sales price levels and confusion matrices for the overall estimation performances of the constructed Bayesian network model were obtained for both the training and the testing data in the following sections.

(a) Model validation results for the training data

The training data consist of 3500 observations. The AUC in Fig. 3 shows that the Bayesian network model estimated the sales prices of the flats that are between 0 and 200,000 TL with a correct rate of 91%.

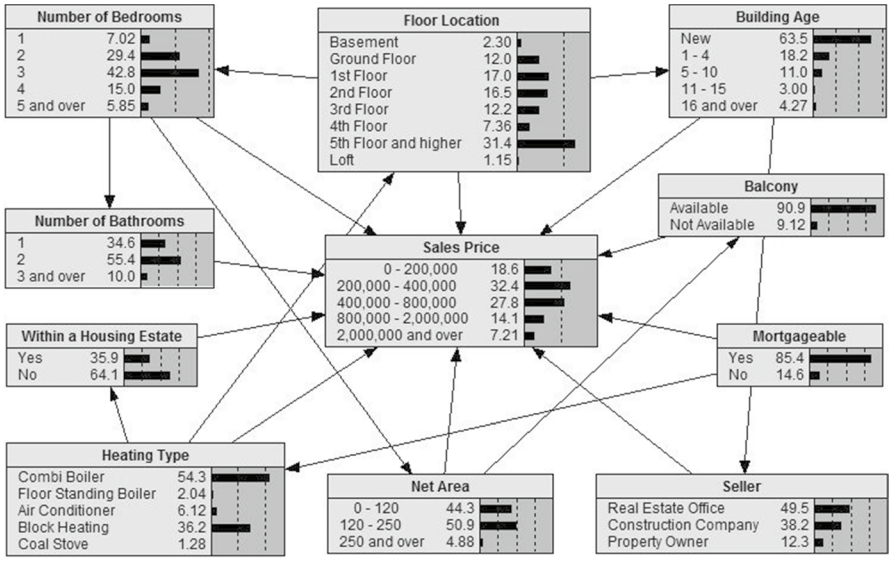

Fig. 2 Bayesian network for estimating the sales prices of flats depending on their characteristics

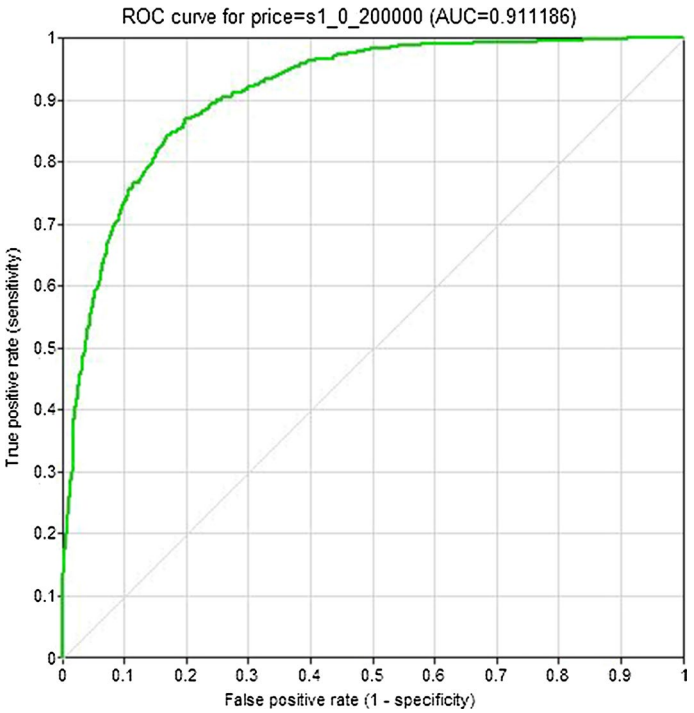

Fig. 3 Training data ROC curve for the sales prices 0–200,000 TL

Similarly, The AUC score in Fig. 4 shows that the Bayesian network model estimated the sales prices of the flats that are between 200,000 and 400,000 TL with a correct rate of 80%.

For flats having sales prices between 400,000 and 800,000 TL the AUC score indicated a correct estimation rate of approximately 82% as seen in Fig. 5.

In Fig. 6, the correct estimation rate is approximately 86% for flats having sales prices between 800,000 and 2,000,000 TL as indicated by the AUC score.

Finally, speaking of flats having sales prices 2,000,000 TL and over, the correct estimation rate of the Bayesian network model was given by the AUC score, which is provided in Fig. 7, as approximately 96%.

The AUC scores provided by the training data are briefly demonstrated in Table 3.

The confusion matrix given in Table 4 shows the correct classification success of the model estimated by the training data. The diagonal values of the matrix indicate the correctly classified values among the predicted and actual sales prices.

In Table 4, the proportion of the correctly classified values can be used to calculate the overall accuracy rate of the model for all the flat sales price levels. This proportion indicates an overall accuracy level for the model as approximately 60%, when the training data set is used.

(b) Model validation results for the testing data

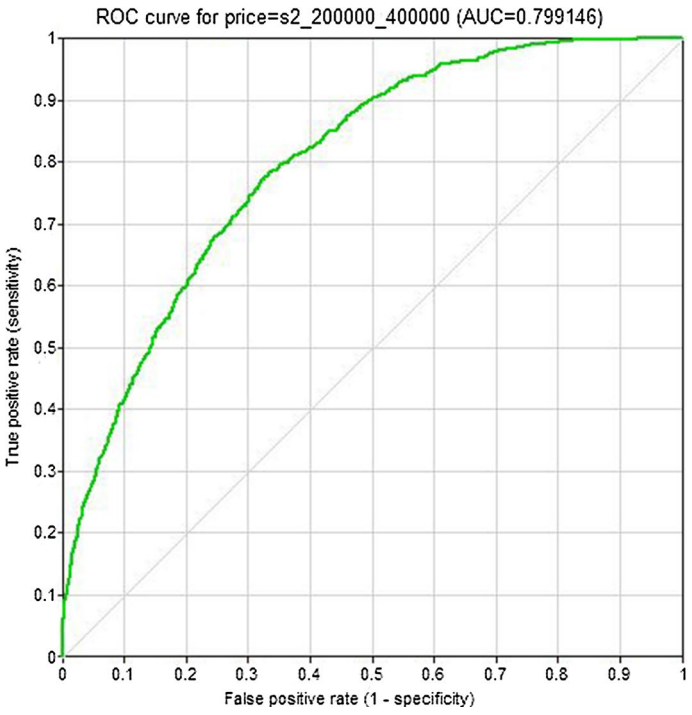

Fig. 4 Training data ROC curve for the sales prices 200,000–400,000 TL

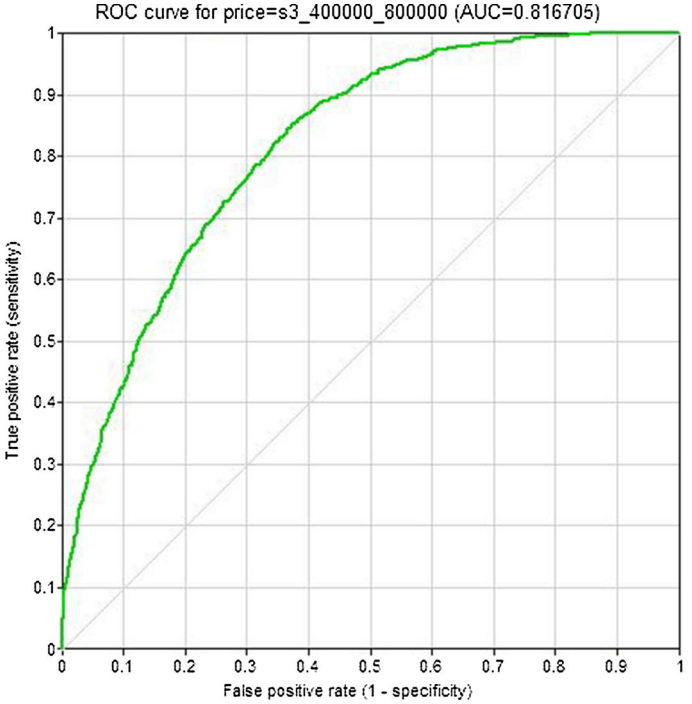

Fig. 5 Training data ROC curve for the sales prices 400,000–800,000 TL

When the estimated Bayesian network model was tested with the help of the testing data, the corresponding AUC scores and the confusion matrix are given below.

The AUC in Fig. 8 shows that based on the testing data, the Bayesian network model estimated the sales prices of the flats that are between 0 and 200,000 TL with a correct rate of 90%.

Similarly, The AUC score in Fig. 9 shows that using the testing data, the Bayesian network model estimated the sales prices of the flats that are between 200,000 and 400,000 TL with a correct rate of 81%.

When the testing data is used, for flats having sales prices between 400,000 and 800,000 TL the AUC score indicated a correct estimation rate of approximately 83% as seen in Fig. 10.

In Fig. 11, the correct estimation rate of the model for the testing data is approximately 86% for flats having sales prices between 800,000 and 2,000,000 TL as indicated by the AUC score.

Finally, speaking of flats having sales prices 2,000,000 TL and over, the correct estimation rate of the Bayesian network model for the testing data was given by the AUC score, which is provided in Fig. 12, as approximately 94%.

The AUC scores provided by the testing data are briefly presented in Table 5.

The confusion matrix given in Table 6, shows the correct classification success of the model when the testing data is used. The diagonal values of the matrix indicate the correctly classified values among the predicted and actual sales prices.

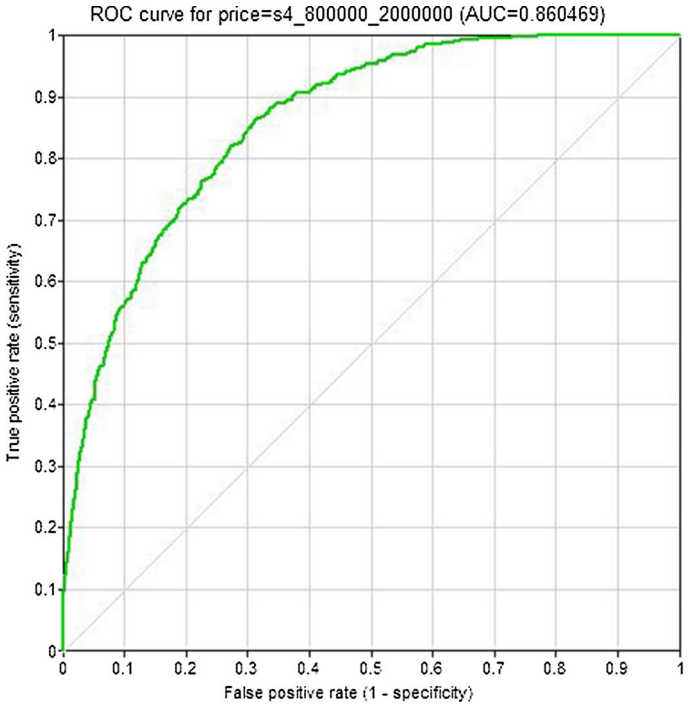

Fig. 6 Training data ROC curve for the sales prices 800,000–2,000,000 TL

In Table 6, the proportion of the correctly classified values indicates an overall accuracy rate for the model as 61% for all levels of the flat sales prices, when the testing data is used.

When the performance of the model with the training data (60%) and the testing data (61%) are compared, it is possible to conclude that the estimation performance of the model with the testing data is slightly better than the one with the training data. However, both 60% and 61% accuracy rates can be accepted to be satisfactory for such a big data set having 3500 cases and 1500 cases for the training data and the testing data respectively with 11 nodes and 43 levels of them in total.

In order to compare the estimation performance of the constructed Bayesian network model with the performances of some other machine learning algorithms, an Artificial Neural Network (ANN) model using multilayer perceptron function, a Decision Tree (DT) model and a logistic model were also constructed employing Weka (2021) software. During the construction of each model, the data set was split into two parts as 70% training set and 30% testing set. The confusion matrix of the ANN model for the flat sales prices is given in Table 7.

When Table 7 is examined the correct classification rate of the ANN model for the sales prices is calculated as 49.8%. The confusion matrix of the DT model for the flat sales prices is given in Table 8.

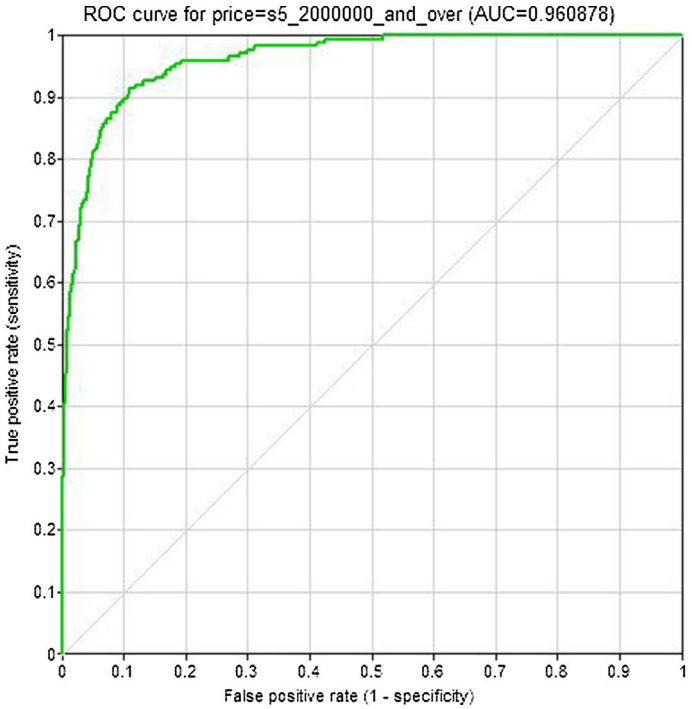

Fig. 7 Training data ROC curve for the sales prices 2,000,000 TL and over

Table 3 AUC scores for the training data

Flat sales prices (TL)	AUC scores (training data)
0–200,000	0.911186
200,000–400,000	0.799146
400,000–800,000	0.816705
800,000–2,000,000	0.860469
2,000,000 and over	0.960878

In Table 8, it is seen that the correct classification rate of the DT model for the sales prices is calculated as 48.5%. The confusion matrix of the logistic model for the flat sales prices is given in Table 9.

Table 9 shows that the correct classification rate of the logistic model for the sales prices is calculated as 49.3%. Table 10 summarizes the correct classification rates of the models.

The results in Table 10 show that The Bayesian network estimated the flat sales prices with the highest performance level of 61% and this level is significantly higher than the levels of the other models. The next best score belongs to the ANN model as 49.80%. The third successful estimation rate is 49.30% that

Table 4 Confusion matrix of flat sales prices for the training data

Flat sales prices (TL)	Predicted				
	0–200,000	200,000 –400,000	400,000 –800,000	800,000 –2,000,000	2,000,000 and over
Actual					
0–200,000	399	195	41	10	3
200,000–400,000	145	725	232	40	4
400,000–800,000	22	260	670	70	26
800,000–2,000,000	10	61	195	182	33
2,000,000 and over	0	8	32	25	112

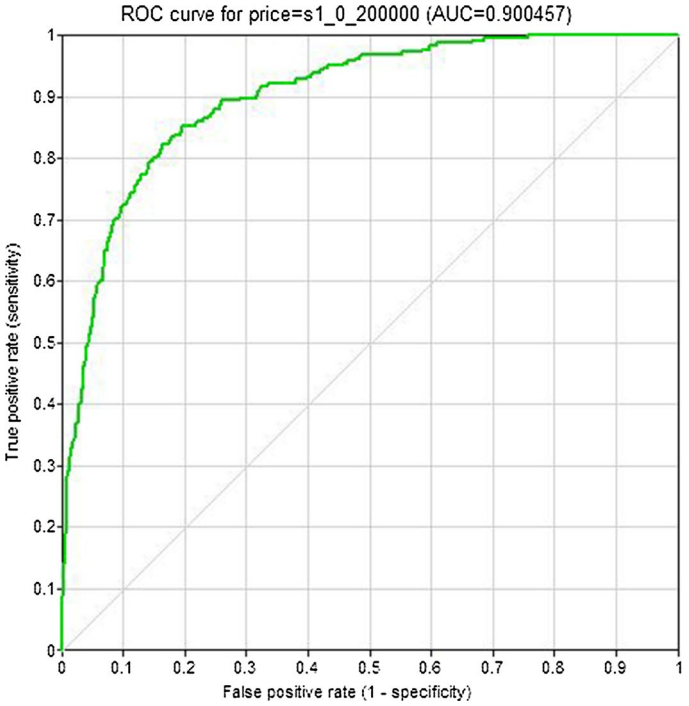

Fig. 8 Testing data ROC curve for the sales prices 0–200,000 TL

belongs to the logistic model and the model having the least successful performance level is the DT model with a score of 48.5%. The comparison can also be made by concluding that the Bayesian network, which is a graphical model, performed the best in this study. The ANN model and the logistic model, which are additive types of models, performed significantly less than the Bayesian network. Finally, the DT, which is a nonparametric machine learning algorithm, gave the lowest estimation performance.

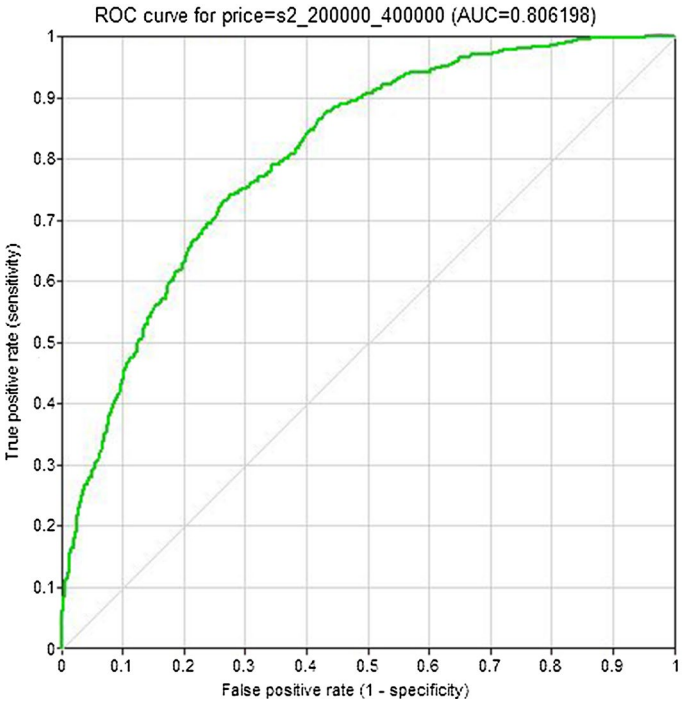

Fig. 9 Testing data ROC curve for the sales prices 200,000–400,000 TL

5.3 Sensitivity Analysis

It is possible to see the degree of influence of other variables on a given target variable by a sensitivity analysis performed for a Bayesian network. Entropy score can be used for sensitivity analysis. The entropy score is calculated as follows (Wiegerinck and Heskes 2001).

$$H(p) = -\sum_{\{x\}} P(x)\log P(x) \tag{4}$$

The entropy scores that measures how much the sales price variable is affected by the changes in the levels of other variables are given in a descending order in Table 11. The results were obtained using Netica (2019) software.

When the sensitivity analysis results in Table 11 are examined, it is seen that the sales price of flat type residence is mostly affected by the number of bathrooms in the flat. Thus, it can be concluded that the most effective factor changing the sales price of a flat is the number of bathrooms. As the number of bathrooms increases, the price of the flat appears to increase quite significantly.

The second most important factor affecting the sales price is the number of bedrooms the flat has. According to this result, it can be stated that as the number of bedrooms increase, the sales price of a flat increases considerably.

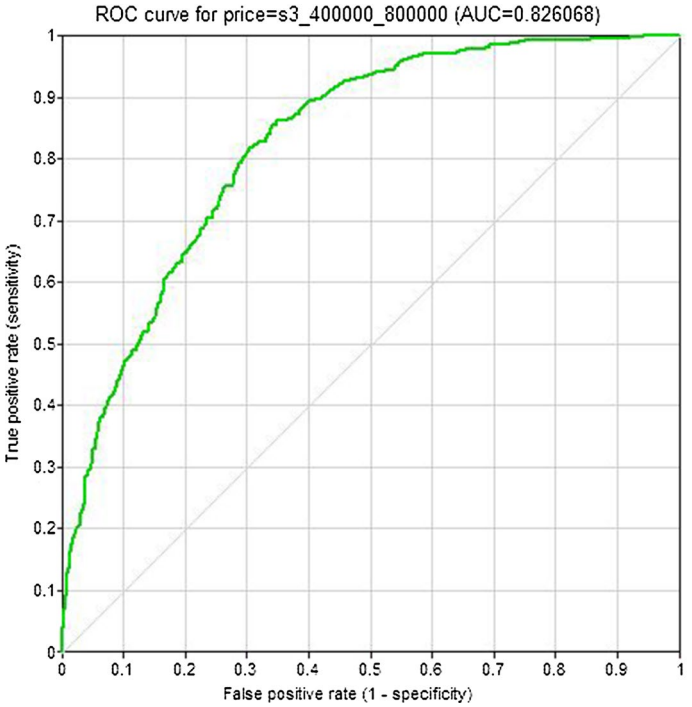

Fig. 10 Testing data ROC curve for the sales prices 400,000–800,000 TL

The third most important price determinant is size which is expressed by the net area of the flat. An increase in the usage area is also an effective reason for the increase in the sales price of flats.

The next important factors affecting the sales price of flats are floor location, heating type and building age respectively. However, suitability for housing estate loan, whether located in a housing estate and availability of balcony do not lead to dramatic changes in the price of a flat. It is also seen that there is not a significant difference among the sales prices of flats being sold by a housing estate office, a construction company or a property owner.

6 Discussions

With the help of the estimated Bayesian network, some findings of the analysis study with respect to 10 characteristics of a flat are given as follows.

When the flats with a sales price of TL 2,000,000 or more are taken into consideration, it is seen that 52.6% of these residences are newly built flats and 63.3% of them are not in a housing estate area. 52.5% of these flats are being sold by real estate offices, 31.1 of them are being sold by construction companies and 16.4% are being sold by property owners. 85.1% of these flats have balconies and 34.1% are located on the 5th or higher floors. 51% of them have a net area of 120–250 m^2 and

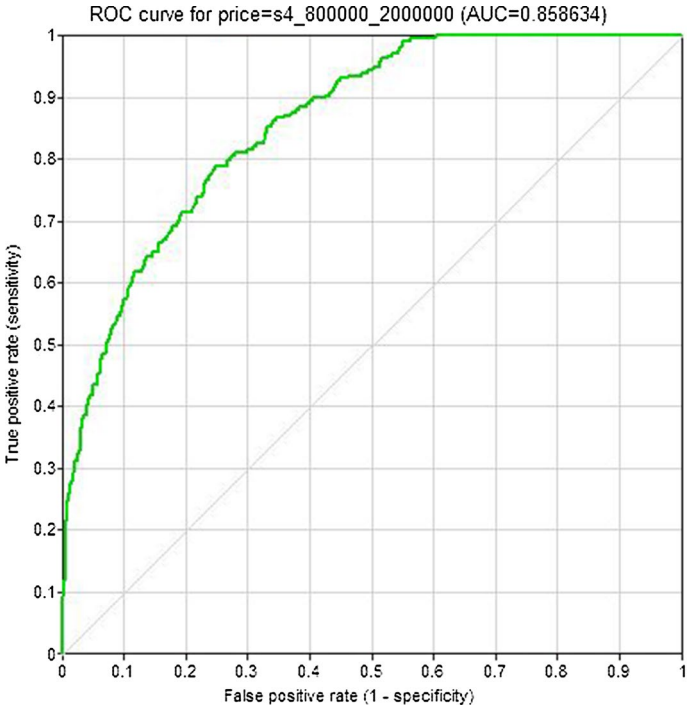

Fig. 11 Testing data ROC curve for the sales prices 800,000–2,000,000 TL

19.7% of them have a net usage area of 250 m^2 or more. 44.2% of these flats, which are in the most expensive category, have block heating, 40% have combi boiler. Moreover, 75.3% of them are suitable for mortgage credit, 25.5% have 4 bedrooms and 20% have 5 or more bedrooms.

One of the common characteristics of the flats in the lowest sales price category is that they are located mostly on the first floor or on the ground floor (22.7% and 19.6%). More than half of these types of flats (51%) are being sold by real estate offices and 90% of them have balconies. 70.7% of them have a net usage area between 0 and 120 m^2. Majorities (67.1%) of the flats with the lowest sales prices use combi boilers for heating. 84.4% of them are suitable for mortgage credit and 45.4% of them have 2 bedrooms. The proportions of these types of flats having 3 bedrooms and 1 bedroom are 27.9% and 17.8% respectively. Moreover, the vast majorities (63.9%) of these flats have a single bathroom and 65.1% are newly built.

As far as the effects of the number of bathrooms on flat sales prices are considered, among the flats with 3 or more bathrooms, 23.7% have sales prices of 2,000,000 TL or more, 27.8% of them have a price between 800,000 and 2,000,000 TL and 24.7% have a price between 400,000 and 800,000 TL. If there are 2 bathrooms in flats, the sales prices are between 400,000 and 800,000 with a probability of 35.3%, between 200,000 and 400, 000 with a probability of 32.5% and 800,000–2,000,000 TL with a probability of 16.2%. Majorities of flats with a single bathroom have sales prices between 200,000 and 400,000 and 0–200,000 with

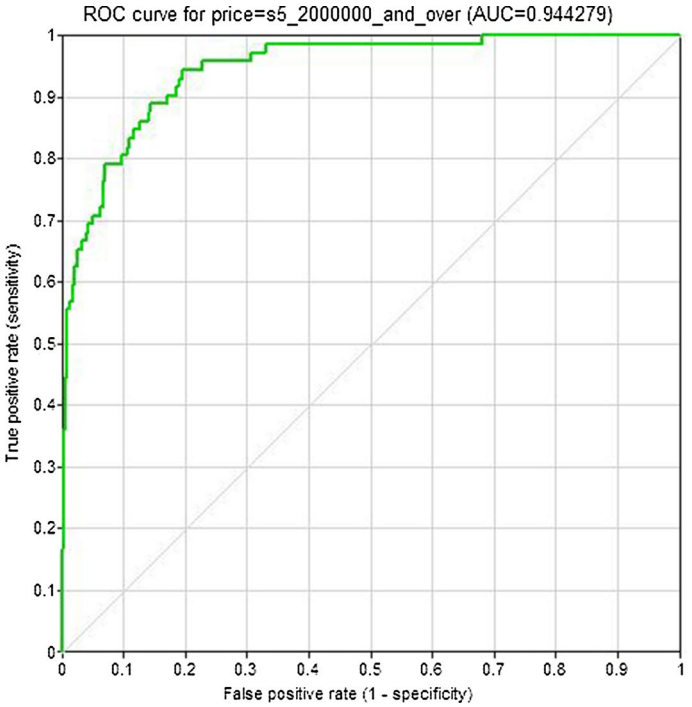

Fig. 12 Testing data ROC curve for the sales prices 2,000,000 TL and over

Table 5 AUC scores for the testing data

Flat sales prices (TL)	AUC scores (testing data)
0–200,000	0.900457
200,000–400,000	0.806198
400,000–800,000	0.826068
800,000–2,000,000	0.858634
2,000,000 and over	0.944279

probabilities 37.5% and 34.3% respectively. Results show that, as the number of bathrooms increases, the sales prices significantly increase.

Considering the number of bedrooms available in flats, those having 1 bedroom are sold in the range of 0–200,000 TL with a probability of 47.2%. The probability that they are traded in the range of 200,000–400,000 is 28.4%. The most popular price range for flats having 2 bedrooms is between 200,000 and 400,000 TL with a probability of 36.5%, while flats with 3 bedrooms are sold in the ranges of 200,000–400,000 TL and 400,000–800,000 TL with probabilities of 36.5% and 33.3% respectively. Flats having 4 bedrooms are sold in the 400,000–800,000 TL band with the highest probability of 34.7%, but the second

Table 6 Confusion matrix of flat sales prices for the testing data

Flat sales prices (TL)	Predicted				
	0–200,000	200,000–400,000	400,000–800,000	800,000–2,000,000	2,000,000 and over
Actual					
0–200,000	169	71	21	4	1
200,000–400,000	71	327	92	16	4
400,000–800,000	12	100	282	34	6
800,000–2,000,000	3	34	81	88	12
2,000,000 and over	0	5	17	8	42

Table 7 Confusion matrix of the Artificial Neural Network model for the flat sales prices

Flat sales prices (TL)	Predicted				
	0–200,000	200,000–400,000	400,000–800,000	800,000–2,000,000	2,000,000 and over
Actual					
0–200,000	148	83	24	4	1
200,000–400,000	76	274	138	24	1
400,000–800,000	15	134	248	31	6
800,000–2,000,000	11	41	89	44	33
2,000,000 and over	1	10	19	12	33

Table 8 Confusion matrix of the decision tree model for the flat sales prices

Flat sales prices (TL)	Predicted				
	0–200,000	200,000–400,000	400,000–800,000	800,000–2,000,000	2,000,000 and over
Actual					
0–200,000	161	75	15	9	0
200,000–400,000	114	254	116	25	4
400,000–800,000	42	132	214	39	7
800,000–2,000,000	10	41	84	70	13
2,000,000 and over	1	9	15	22	28

intensive range is realized as 800,000–2,000,000 TL with a probability of 23.8%. Flats with 5 and more bedrooms are offered for sale at prices between 800,000 and 2,000,000 TL with a probability of 32.6% and 2,000,000 TL and more with a probability of 24.7%.

Table 9 Confusion matrix of the logistic model for the flat sales prices

Flat sales prices (TL)	Predicted				
	0–200,000	200,000–400,000	400,000–800,000	800,000–2,000,000	2,000,000 and over
Actual					
0–200,000	147	99	13	0	1
200,000–400,000	92	263	140	17	1
400,000–800,000	16	135	252	27	4
800,000–2,000,000	8	30	106	50	24
2,000,000 and over	0	4	20	23	28

Table 10 The correct classification rates of all models employed in the study

	The Bayesian network model	The artificial neural network model	The decision tree model	The logistic model
Number of correctly classified cases	908	747	727	740
Number of incorrectly classified cases	592	753	773	760
Percentage(%)	61	49.80	48.50	49.30

Table 11 Sensitivity analysis for flat sales prices

Variable	Entropy score
Number of bathrooms	0.12302
Number of bedrooms	0.11548
Net area	0.07552
Floor location	0.03940
Heating type	0.03678
Building age	0.01949
Mortgageable	0.00656
Within a housing estate	0.00610
Seller	0.00450
Balcony	0.00304

When the net usage areas of the flats are taken into consideration, as expected, the sales prices increase as the areas increase. When the ages of the flats are considered, however, the sales prices decrease as the ages of buildings increase.

On the other hand, if the ages and net areas of the flats are considered together, flats which are 16 years or older with a net area of 250 m^2 or over are sold for 2,000,000 TL and above with the highest probability of 21.4% or for

800,000–2,000,000 TL with the second highest probability of 20.6%. Moreover, newly-built flats with a net usage area of 120–250 m^2 are offered for sale in a range of 200,000–400,000 TL with a probability of 36.5% and between 400,000 and 800,000 TL with a probability of 35%. According to these results, it is seen that sales prices are not affected by ages of buildings as long as the usage areas of the flats are large. In other words, prices of flats do not decrease dramatically in time, as long as they have large usage areas.

Price estimation of dwellings has been made using conventional hedonic pricing models for long. Sirmans et al. (2005) examined more than 125 empirical hedonic pricing model studies and they found out that the studies often disagree on the effects of certain characteristics on pricing. According to Sirmans et al. (2005) number of bathrooms, number of bedrooms, net area and age variables are among the twenty characteristics appearing most often in hedonic pricing model studies. There are also other studies involving some of the chosen factors in our study and providing some results related to them.

Considering number of bathrooms, Sirmans et al. (2005) reported that among 40 studies they examined, 34 of them reported a positive relation between the number of bathrooms in a house and its sales price, just 1 of them implied a negative relation and 5 of them found no relation. This result is highly consistent with our finding that number of bedrooms is a factor affecting the sales price of a flat positively, also the most effective one.

Similarly, another result of the Bayesian network model is that number of bedrooms significantly and positively affects the sales price of a flat. Sirmans et al. (2005) report that among 40 hedonic pricing model studies they examined, 21 studies found that number bedrooms have positive impact on house price, 9 of them indicated a negative relation and 10 studies could not detect a significant relation between number of bedrooms and house prices. These results are also supporting the finding of the Bayesian network model about number of bedrooms.

In our study, the finding that the third most effective factor is net area, which affects the flat prices positively, appears to be supported also by the findings of hedonic pricing models as Sirmans et al. (2005) reported that 62 studies out of 69 found that area of a house affects its price positively. Moreover, it affected the price negatively only in 4 studies and no significant relation could be detected in 3 studies.

The findings of the Bayesian network model related to the factors number of bathrooms, number of bedrooms and net area are also supported by another research performed by Caglayan and Arikan (2011) in Istanbul, Turkey. Using a quantile logistic regression model, they found that increasing numbers of bedrooms and bathrooms increase the house prices. Zietz et al. (2008) also reached the same results for the same factors by using a quantile regression model.

Speaking of building age, the result of the Bayesian network, which reports a negative significant effect, is still in accordance with the results provided by Sirmans et al. (2005) who determined that out of 78 studies, 63 of them reported a negative significant effect, 7 of them a positive effect and 8 of them no significant effect between age of buildings and sales prices. Selim (2009) examined the determinants of house prices in Turkey based on the comparison of a hedonic and an artificial neural network model. One of the findings is that the prices of houses between 5 and

10 years of age are less than those that are 0–5 years of age by 5.8% for urban areas. This finding is consistent with the finding in our study that age affects the flat prices but not with great rates.

Existence of balcony did not appear to be a significant factor affecting the flat sales prices in our study. On the contrary, Sirmans et al. (2005), reports that among 12 hedonic pricing model studies, 10 of them reported that balcony factor has a positive effect on house prices, while just 2 of them reported it as an insignificant factor.

According to these findings, while the results of the Bayesian network model related to bathroom, bedroom, net area and age factors are in accordance with the majority group in hedonic pricing models, the result related to balcony factor falls into the minority group. These differences among the results of different studies can be explained by one of the findings of Sirmans et al. (2006) who performed a meta regression analysis with nine housing characteristics and found that the estimated coefficients for some housing characteristics determining house prices vary significantly by geographical location.

As far as heating type factor is concerned, Selim (2009) reports that the prices of the houses heated by block heating and combi boilers are significantly higher than those with stove. The same result was also provided by the Bayesian network model. Still, in our study, heating type has less impact on flat sales prices compared to bathroom, bedroom and area factors.

Floor location variable, which is another factor having a comparatively low impact rate on flat prices in our study, showed that prices of the flats located in the basement or on the ground and 1st floors are slightly less than the prices of those on the higher floors. The finding of our study is supported by one of the findings in another study performed in Athens, Greece. Efthymiou and Antoniou (2013) compared the hedonic price models and spatial econometric models and they found that the purchase price does not increase linearly but continuously with the floor level.

One of the results provided by the Bayesian network is that mortgageability of a flat does not seem to have a significant effect on its sales price. This result can be explained by observing the levels of mortgageable node in the Bayesian network model given in Fig. 2 which indicates that the probability that a flat in Turkey is eligible for mortgage credit is considerably high (85.4%).

Dekker and van Kempen (2005) provide an analysis of large housing estates in Europe. They suggest that the prices of the housing units are mostly cheap and affordable range, yet the differences between countries can be huge. Additionally, they state that all the estates are relatively cheap places to live but necessarily the cheapest in the city. These comments are consistent with the results of the Bayesian network model in our study which suggests that a flat being in a housing estate does not have a significance effect on its sales price. In Turkey, housing estates are common especially in big cities. However, they cannot be seen as homogeneous housing units as there could be significant differences among the characteristics of them.

When the impact of the seller type on flat prices is considered, our study shows that there is no significant difference among the prices of the flats being sold by a real estate office, a construction company or a property owner. On the other hand, in a study in the USA given by Jauregui and Hite (2010) shows that real estate agents obtain higher prices than expected, when the houses are located closer to an

environmental disamenity. They attribute this result to the bargaining power of the real estate agents on the sales prices to increase their commissions. The difference between our results and theirs can be explained by the environment we sampled from. We have derived the data from a property website where real estate offices, construction companies or property owners advertise under the same conditions and rules. Visitors are able to make a search with respect to various attributes of the properties including sales prices. Thus, real estate agents or the other advertisers on the website must consider the equilibrium of prices in the environment, especially if they want to sell their property as fast as possible.

The findings of the Bayesian network model and the hedonic pricing models available in the literature are provided in Table 12.

The detailed comparative results provided in Table 12 are demonstrated in Fig. 13 with brief explanations and consistency percentages where possible.

When Table 12 and Fig. 13 are examined, it can be concluded that except for the two cases, where no comparable data are available belonging to any hedonic price modeling study, the majority of the results of the Bayesian network model are compatible with or parallel to the hedonic pricing model results. The only case, in which the Bayesian network model and the hedonic pricing model result gave completely opposite results, is the estimation of the relation between seller type and sales prices. While, the Bayesian network model suggested no relation between the seller type and flat sales prices, the hedonic pricing model by Jauregui and Hite (2010) determined a significant relation between these two variables.

7 Conclusions

In this study, we have presented a novel approach to estimate flat sales prices based on various characteristics of the flats, as an alternative to the classical hedonic pricing models. A Bayesian network model was estimated based on a machine learning algorithm called PC structural learning algorithm. Bayesian network models are more suitable to be used in residence sales price estimation than hedonic pricing models based on regression. Indeed, the estimated model showed a satisfactory level of accuracy and provided results highly in accordance with more than 125 hedonic pricing models in the literature that Sirmans et al. (2005) examined. For example, in hedonic price modeling studies, number of bathrooms, number of bedrooms, net area and age variables are found to be significant characteristics affecting the sales prices of residences and these variables also appeared to be significant factors in our Bayesian network model. However, the Bayesian model estimated has an important advantage over the hedonic price models such that as any variable in the network changes, not only the dependent variable changes but all the related variables in the network changes too. For instance, using the Bayesian network price estimation model, when buyers decide about the floor location and the number of bedrooms in the flat they intend to buy, they will not only see the most probable prices that they will pay, but also the probabilities related to the heating types they will have and the probabilities related to the age of the building they will live in. Future work includes

Table 12 Findings of the Bayesian network model and the hedonic pricing models

	The Bayesian network model	The hedonic pricing models
Number of bathrooms and sales prices		42 models were examined
	Reports a positive relation	36 models report a positive relation
		1 model reports a negative relation
		5 models report no relation
Number of bedrooms and sales prices		42 models were examined
	Reports a positive relation	23 models report a positive relation
		9 models report a negative relation
		10 models report no relation
Net area and sales prices		69 models were examined
	Reports a positive relation	62 models report a positive relation
		4 models report a negative relation
		3 models report no relation
Building age and sales prices		79 models were examined
		7 models report a positive relation
	Reports a negative relation	64 models report a negative relation
		8 models report no relation
Existence of balcony and sales prices		12 models were examined
		10 models report a positive relation
		0 model reports a negative relation
	Reports no relation	2 models report no relation
Heating type and sales prices		1 model was examined
	Reports a strong significant effect	Reports a strong significant effect
Floor location and sales prices		1 model was examined
	Reports a weak significant effect	Reports a continuous relation between them instead of linear
Mortgageability and sales prices		
	Reports no significant effect	No record
Being in a housing estate and sales prices		
	Reports no significant effect	No record
Seller type and sales prices		1 model was examined
	Reports no significant effect	Reports a significant effect

Fig. 13 Brief comparisons between the Bayesian network model and the hedonic pricing models with the consistency percentages

re-building the Bayesian network model with different characteristic variables or building it for different countries by the addition of some country-specific residence characteristics. The model could also be adapted as a smart device application to be used in property sales markets.

Author Contributions All work done on this article belongs to the sole author of the article.

Funding No funding was received from any resource.

Availability of Data and Material (Data Transparency) The data could be shared on a reasonable request.

Code Availability (Software Application or Custom Code) GeNIe (2019) (free for academic use), Netica (2019) (freeware version) and Weka (2021) (freeware) software packages have been employed and they are cited in the references section.

Compliance with Ethical Standards

Conflict of interest The author declares that he has no conflict of interest.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & B. F. Csaki (Eds.), *Second International Symposium on Information Theory*, Academiai Kiado: Budapest.

Caglayan, E., & Arikan, E. (2011). Determinants of house prices in Istanbul: A quantile regression approach. *Quality & Quantity, 45*(2), 305–317.

Dekker, K., & van Kempen, R. (2005). Large housing estates in Europe: A contemporary overview. In R. van Kempen, K. Dekker, S. Hall, & I. Tosics (Eds.), *Restructuring Large Housing Estates in Europe* (pp. 19–46).

Efthymiou, D., & Antoniou, C. (2013). How do transport infrastructure and policies affect house prices and rents? Evidence from Athens, Greece. *Transportation Research Part A: Policy and Practice, 52,* 1–22.

Eurostat (2018). Retrieved September 22, 2020, from https://ec.europa.eu/eurostat/statisticsexplained/index.php/Living_conditions_in_Europe_housing_quality.

GeNIe (2019). BayesFusion, LLC. Retrieved October 22, 2020, from http://www.bayesfusion.com/.

Giudice, V. D., Paola, P. D., Forte, F., & Manganelli, B. (2017). Housing estate appraisals with bayesian approach and markov chain hybrid monte carlo method: An application to a central urban area of naples. *Sustainability,* 9(11).

Heckerman, D., Chickering, D. M., Meek, C., Rounthwaite, R., & Kadie, C. (2000). Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research, 1,* 49–75.

Holmes, D. E., & Jain, L. C. (2008). *Innovations in Bayesian networks: Theory and applications*. Berlin: Springer.

Hui, S. K., Cheung, A., & Pang, J. (2010). A hierarchical bayesian approach for residential property valuation: Application to Hong Kong housing market. *International Housing estate Review, Asian Housing estate Society, 13*(1), 1–29.

Jauregui, A., & Hite, D. (2010). The impact of real estate agents on house prices near environmental disamenities. *Housing Policy Debate, 20*(2), 295–316.

Jensen, F. V., & Nielsen, T. D. (2007). *Bayesian networks and decision graphs*. New York: Springer.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy, 74*(2), 132–157.

Liu, Z., Yan, S., Cao, J., Jin, T., Tang, J., Yang, J., & Wang, Q. (2018). A Bayesian approach to residential property valuation based on built environment and house characteristics. In *2018 IEEE International Conference on Big Data*, Seattle, WA, USA.

Meek, C. (2013). Strong completeness and faithfulness in Bayesian networks. Preprint http://arxiv.org/abs/1302.4973.

Netica (2019). Norsys Software Corp. Retrieved October 22, 2019, from https://www.norsys.com/netica.html.

Pearl, J. (1985). Bayesian networks: A model of self–activated memory for evidential reasoning. In *7th Conference of the Cognitive Science Society*, California, USA.
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.
Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications, 36*(2), 2843–2852.
Sirmans, G. S., MacDonald, L., Macpherson, D. A., & Zietz, E. N. (2006). The value of housing characteristics: A meta-analysis. *The Journal of Real Estate Finance and Economics, 33*(3), 215–240.
Sirmans, G. S., Macpherson, D. A., & Zietz, E. N. (2005). The composition of hedonic pricing models. *Journal of Real Estate Literature, 13*(1), 3–46.
Spirtes, P., & Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review, 9*(1), 67–72.
Spirtes, P. (2010). Introduction to causal inference. *Journal of Machine Learning Research, 11*(May), 1643–1662.
Tsagris, M. (2019). Bayesian network learning with the PC algorithm: an improved and correct variation. *Applied Artificial Intelligence, 33*(2), 101–123.
TurkStat (2011). Turkiye Nufus ve Konut Arastirmasi. Retrieved October 22, 2019, from www.tuik.gov.tr.
Verma, T., & Pearl, J. (1990). Equivalence and synthesis of causal models, Sixth Annual Conference on Uncertainty in Artificial Intelligence, Massachussets, USA, 255–270, July, 27–29.
Weka (2021), Retrieved January 19, 2021, from https://www.cs.waikato.ac.nz/ml/weka/.
Wiegerinck, W. A. J. J., & Heskes, T. (2001). Probability assessment with maximum entropy in Bayesian networks. *Computing Science and Statistics, 33,* 1–9.
Zietz, J., Zietz, E. N., & Sirmans, G. S. (2008). Determinants of house prices: A quantile regression approach. *The Journal of Real Estate Finance and Economics, 37*(4), 317–333.