CONDUCTING THE BEST HOUSE PRICE PREDICTION MODEL FOR TURKISH HOUSING MARKET WITH MACHINE LEARNING

[1]DUYGU DORUKHAN, [2]HILAL ERDOGAN SUMNU, [3]HAKAN TAS, [4]KAMIL CAYIRLI

[1]Specialist at Data Science, Zingat.Com, Istanbul/ Turkey
[2]Head of Data Management, Zingat.Com, Istanbul/ Turkey
[3]Head of Data Science, Zingat.com, Istanbul/ Turkey
[4]Lecturer of nişantaşıüniversitesi, Istanbul/ Turkey
E-mail: [1]duygu.dorukhan@zingat.com, [2] hilal.sumnu@zingat.com, [3] hakan.tas@zingat.com, [4]kamil.cayirli@nisantasi.edu.tr

Abstract - The objective of this paper is to empirically conduct the best machine learning regression model for Turkish Housing Market by comparing accuracy scores and absolute deviations of test results. This paper uses the locational listing data of Turkish leading property portal Zingat.com, for the four-and-a-half-year period from January 2015 to June 2019. In our study, fifteen explanatory building and dwelling variables are used for each prediction model; Location (latitude& longitude), Dwelling Size, Terrace Size, Number of Room, Number of Bathroom, Building Age, Total Floor, Floor Number, Property Type, View, View Direction, Lift, Heating System, Security, and Parking. Data preprocessing (cleaning contradictory data, completing missing data etc.) was done on dataset. Dataset was separated as rental listings and for sale listings. Training the data model; data was apportioned into training and test sets, with an 80-20 split. Three different data models were created by using Support vector machine, Feedforward Neural Networks, Generalized Regression Neural Networks algorithms of forecasting models with supervised learning. Python programming language and Keras library were used. Mean Absolute Percentage Error (MAPE) was selected as the error metric. When the success of the methods was examined according to the results of the analysis, it has been found that the Feedforward Neural Network model gives better results than the other supervising learning models. Smallest MAPE observed as 9.3% for rental listings and 14.9% for on sale listings. The most important variable in the model was building location when we consider the importance of the 15 variables in network development. Dwelling size was the 2nd the most important variable and terrace size had the least importance.

Keywords - Turkey, Housing Market, Machine Learning, Zingat.com, House Price Prediction.

I. INTRODUCTION

In Turkey, housing is one of the important and traditional investment for households. However, housing is a relatively illiquid investment, with an uncertain capital value, and it is generally highly leveraged, which makes it a potentially important channel of transmission of monetary policy. Due to population growth, decrease in average size of the household, migration and rapid pace of urbanization, housing demand is continuously increasing in Turkey, and since this requirement cannot be met with supply, house prices have been increasing steadily. To monitor the evolution of house prices and the market trend over time, related data is gathered from Turkish leading property portal Zingat.com.
Zingat.com is meeting point and source of informationfor investors in the real estate sector,real estate sellers / buyers, people who is looking for a new house, professionals who earn in this sector.In this respect, it aims to provide accurate real estate appraisal services to its users.The objective of this paper is to empirically conduct the best machine learning regression model for Turkish Housing Market.

II. LITERATURE REVIEW

The relationship between machine learning and the real estate market has come to prominence in recent years, especially with the development of technology and studies on data analysis. Nilsson (2019) tried to estimate the selling prices of properties in Stockholm using the parameters of apartmentsin artificial neural networks. Networks were trained both for pricing based on the list price and to give the apartment a new price. Phan (2018) has applied machine learning techniques to analyze the purchase and sale of historic properties in Australia to find useful models for home buyers and sellers. The result of the study was the high inconsistency between the most expensive and affordable home prices in Melbourne. Furthermore, experiments have shown that the combination of Stepwise and Support Vector Machines based on the mean square error measurement is a competitive approach. Ćetković et all. (2018) aimed to establish a prognostic model of real estate market value in EU countries due to the impact of macroeconomic indicators. The available inputs have shown that macroeconomic variables affect the determination of property prices. Varma et all. (2018) used various regression techniques and the results were calculated as the weighted average of various techniques. The results show that this approach provides minimum error and maximum accuracy compared to the individual algorithms applied. Zhou et all. (2018) proposes the Genetic Algorithm, Geographical Information, Support Vector Machine model for the estimation of housing prices and combines artificial neural networks with Particle Swarm Optimization. In this study, we focus on the current literature on artificial neural networks.

Finally, proposals were made for the assessment of housing in China. New Support Vector Machine (SVM) approach was applied for the estimation of house prices byChen et all. (2018). The results calculated with non-parametric estimates confirm that the predictive power of using SVM in the house price estimation is highly accurate. In addition, studies to extract geographic weights using core density estimates to reflect price responses to quantities of local hedonic properties are recommended. Abidoye and Chan (2018) use Artificial Neural Network (ANN) in real estate appraisal using the Lagos metropolitan real estate market as a representative case. The results show that the ANN model has a good estimation ability, is suitable and reliable for property valuation and can be used as a tool by the housing market, especially valuation and researchers for pricing.

III. THEORETICAL FRAMEWORK AND DATA

Support Vector Machines, Feedforward Artificial Neural Networks and Generalized Regression Neural Networks will be used in the development of real estate valuation estimation models. Performance of the support vector machines and performance of the housing valuation estimation models to be developed are directly proportional to the determination of the optimal values of the kernel function and the parameters of the function. In this study, Radial, Linear, Polynomial and Sigmoid functions were tested as kernel functions. The parameter values of the specified kernel functions and the optimum values of the Epsilon and cost parameters of the Support Vector Machines were determined by grid search. The performance of real estate valuation estimation models developed with Feedforward Neural Networks depends on many parameters such as number of hidden layers to be used in the network, number of neurons to be found in hidden layers, number of epokes, learning coefficient, momentum and activation functions to be used in hidden / exit layers. While estimation models were formed, optimum values of these parameters were found during the training phase. The training process on the dataset continued until a certain number of epokes were reached or below a predetermined threshold error value. The performance of real estate valuation estimation models to be developed with Generalized Regression Neural Networks is used as kernel function. It depends on the parameter value of the Gaussian function. The optimum value of this parameter was determined during the training phase of the data set and neuron pruning was also applied to improve the performance of this method.

Due to the large number of estimation variables affecting the property price, whether more accurate estimates can be obtained with fewer estimation variables by using Relief-F and Minimum Redundancy Maximum Relevance (mRMR) methods was investigated. Mean Absolute Percentage Error (MAPE), which is frequently used in the literature, was used to evaluate the performance of estimation models. Python programming language and Keras library were used to construct prediction models.

Depending on the analysis processes presented above, firstly, data was collected from Zingat.com database for sale / rent houses belonging to Ankara, Istanbul and İzmir and data sets were created. In the data sets, the internal (number of rooms, net area, gross area, number of toilets / bathrooms, number of balconies, heating status, age of building, number of floors, security, parking, facade, current floor, elevator, site, pool) and external variables (location, price per square meter, proximity (school, public transport, shopping center, health institution, etc.), landscape). Preprocessing operations (clearing of outliers, missing data completion, etc.) were performed on the data sets and the data set was divided into two as sale listings and rental listings. The following tables (Table 1 and Table 2) contain detailed data sets.

id	City	County	District	Status	Price (TL)
1	İstanbul	Beylikdüzü	Merkez	Sales	320.000
2	Ankara	Çankaya	İlkbahar	Sales	365.696
3	Ankara	Çankaya	Çankaya	Sales	256.292
4	İstanbul	Beylikdüzü	Dereağzı	Sales	293.893
5	İstanbul	Beylikdüzü	Marmara	Sales	315.200
6	İstanbul	Beylikdüzü	Yakuplu	Sales	400.000
7	İstanbul	Beylikdüzü	Marmara	Sales	280.000
8	İstanbul	Beylikdüzü	Adnan Kahveci	Sales	300.653
9	İstanbul	Beylikdüzü	Büyükşehir	Sales	272.000
10	İstanbul	Beylikdüzü	Kavaklı	Sales	345.000
11	İstanbul	Beylikdüzü	Gürpınar	Sales	300.653
12	İstanbul	Beylikdüzü	Cumhuriyet	Sales	171.523
13	Ankara	Çankaya	Oran	Sales	490.668
14	İzmir	Beylikdüzü	Kavaklı	Sales	265.955
15	Ankara	Çankaya	Ayrancı	Sales	360.540

Table 1. Housing Prices for Sale (2015-2019 Average Price for 100 Sqm House)

id	City	County	District	Status	Price (TL/SQM)
1	Ankara	Çankaya	Oran	Rental	16,50
2	Ankara	Çankaya	Büyükesat	Rental	13,45
3	Ankara	Çankaya	Ayrancı	Rental	9,00
4	Ankara	Çankaya	Ayrancı	Rental	7,99
5	İstanbul	Beylikdüzü	Dereağzı	Rental	6,00
6	Ankara	Çankaya	Sancak	Rental	7,00
7	İzmir	Karşıyaka	Mavişehir	Rental	9,00
8	Ankara	Çankaya	Birlik	Rental	6,60
9	Ankara	Çankaya	Gaziosmanpaşa	Rental	6,50
10	Ankara	Çankaya	Çankaya	Rental	5,50
11	İzmir	Karşıyaka	Yalı	Rental	2,50
12	Ankara	Çankaya	Çankaya	Rental	5,50
13	İstanbul	Beylikdüzü	Barış	Rental	6,50
14	Ankara	Çankaya	Oran	Rental	6,00
15	İzmir	Karşıyaka	Mavişehir	Rental	8,00

Table 2. Housing Prices for Rent (2015-2019 Average Price)

IV. DATA ANALYSIS AND PRESENTATION OF FINDINGS

4.1. House Prices Forecast Model Selection

Training the model; 80% of the data were considered as training set and 20% as test set. 3 different models have been formed with the support vector machines,

feedforward artificial neural networks and generalized regression neural networks. Phyton programming language and Keras library were used in the analysis. Mean Absolute Percentage Error was selected as the error metric. When the success of the methods is examined according to the results of the analysis, it has been found that the artificial neural network model gives better results than the other supervised learning models.

MODEL	GRNN MAPE	MLP MAPE	SVM MAPE
Rental	12,1%	9,3%	11,49%
Sales	19,1%	14,9%	17,93%

Table 3. Model Comparison

4.2. Feedforward Artificial Neural Network

A Multilayer Perceptron (MLP) Network consisting of a total of 34 cells consisting of 1 input, 2 hidden and 1 output layers, 15 in the input layer, 9 in the eachhidden layers and 1 in the output layer was formed. Relu was used in the middle layers of the network and Linear Function was used in the output layer. Epoch number = 1000, Learning Rate = 0.01, Momentum Rate = 0.9 and Optimization function = Adam were used.

MAPE in the study was 9.3% in rental listings; 14.9% for sale listings.

V. CONCLUSION

In this study, we tried to conduct the best machine learning regression model for Turkish Housing Market by comparing accuracy scores and absolute deviations of test results. In our study, fifteen explanatory building and dwelling variables are used for each prediction model; Location (latitude& longitude), Dwelling Size, Terrace Size, Number of Room, Number of Bathroom, Building Age, Total Floor, Floor Number, Property Type, View, View Direction, Lift, Heating System, Security, and Parking. Dataset was separated as rental listings and for sale listings. Training the data model; data was apportioned into training and test sets, with an 80-20 split. Three different data models were created by using Support vector machine, Feedforward Neural Networks, Generalized Regression Neural Networks algorithms of forecasting models with supervised learning. Python programming language and Keras library were used. Mean Absolute Percentage Error (MAPE) was selected as the error metric. In our empirical part, when the 15 input variables of the model were examined, it was observed that the most important variable in the model was location. In the second place is the size of the apartment. The lowest weight belongs to number of balconies. All of these results show that artificial neural network model can be used as an important tool in estimating house prices.

REFERENCES

[1] Pontus Nilsson, Prediction of Residential Real Estate Selling Prices Using Neural Networks, February 2019.

[2] Danh Phan, Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia,2018.

[3] JasminaĆetković et all., Assessment of the Real Estate Market Value in the European Market by Artificial Neural Networks Application, January 2018.

[4] Ayush Varma et all, House Price Prediction Using Machine Learning and Neural Networks,May 2018.

[5] Gang Zhou et all., Artificial Neural Networks and the Mass Appraisal of Real Estate, 2018.

[6] Jieh-Haur Chen et all, Forecasting Spatial Dynamics of the Housing Market Using Support Vector Machine, July 2017.

[7] Rotimi BoluwatifeAbidoye and Albert P. C. Chan, Modelling Property Values in Nigeria Using Artificial Neural Network, February 2014.

★★★