NATIONAL COLLEGE OF IRELAND

THESIS PROPOSAL

Housing Price Forecasting Based on Structural Attributes and Distance to Nearby Schools

Author:
MOHAMMED ABOU HASSAN,
X16150911

M.Sc. in Data Analytics
School of Computing

April 17, 2018

NATIONAL COLLEGE OF IRELAND

Abstract

School of Computing

M.Sc. in Data Analytics

Housing Price Forecasting Based on Structural Attributes and Distance to Nearby Schools

by Mohammed Abou Hassan

Intelligence of machines is increasing with time. Facebook's face recognition, Google's website translation and Siri's voice recognition are perfect examples of what these intelligent systems can do. Powered by statistics and computational research and advancements, machines are discovering insights and predicting unforeseen events. Modern societies depend more and more on these machines across many sectors and housing sector is no exception as it has seen a big increase in reliance on machine learning models built to predict the price of a house in a particular location. Previous studies have demonstrated that taking distance from property to nearby school as a variable can improve regression models. Our work will use this distance in ANN and SVM-R models in addition to structural attibutes and then evaluate the performance of these models using validation techniques which we will discuss thoroughly in later sections. The results of this study can be used to formulate more effective policies for real estate management.

Contents

Chapter 1

Introduction

1.1 Motivation, Research Question and Structure of this Work

Residential properties are considered as one of the most important needs in our modern societies. Buying a residential property is a serious commitment and developing a system that can help buyers pay the right price would be of extreme importance as sellers usually tend to increase the price of a property for profit and knowing the real price could encourage buyers to bid more comfortably (Mukhlishin, Saputra and Wibowo, 2017). Many factors can decide the price of a housing property. Size of rooms, number of bathrooms, surface area of the property and the presence of a garden are the common ones. Many studies have used a combination of these variables to study how the prices of properties vary (Gu, Zhu and Jiang, 2011). Newer studies were successful in adding environmental variables to their models like quality of surrounding neighbourhood, noise and air pollution (Chiarazzo et al., 2014). Recently published papers, highlighted the importance of adding the locations of certain nearby landmarks like train stations or city centres to machine learning models.

Our work will investigate the importance of adding distance to nearby schools to ANN (Artificial Neural Network) and SVM models. This approach has been proved important with regression models (Osland, 2010) and proving it valid for ANN/SVM could open the door to improvements in the way housing prices are predicted. People tend to live near their children's schools to save time and money when dropping them and picking them up which induces the idea of adding distance to nearby schools to the model as an additional feature. Adding such feature should improve the performance of the model and help come up with better predictions. Our research question is formulated as follows: ***"How will ANN and SVM algorithms perform when using distance to nearby schools and structural attributes for housing price prediction?"*** Sections 2.1 and 2.2 will include a literature review that sheds light on previous work, related to our research area (including a critical analysis). Sections 3.1 and 3.2 will provide a thorough explanation about all phases that our work will go through, including data preparation, methodology and evaluation. At the end a Gantt's chart will provide a good idea about the time allocated for each step our research will go through from start to finish.

Chapter 2

Literature Review

A study of the housing price prediction literature reveals the fact that there are two major research trends: the trend of using hedonic regression (statistical method) and the trend of using machine learning algorithms (Park and Kwon Bae, 2015).

2.1 Regression

We start our review by talking about regression models. Although they make accurate predictions and they are widely used, they face a lot of challenges when they deal with outliers, non-linear relations, discontinuity or fuzziness (Park and Kwon Bae, 2015). As a starting point for our research, we discuss the most relevant papers and work done by researchers who were using regression equations to predict property prices. (Brown and Rosen, 1982)'s paper is one of the first papers to be published in that area, where he used a study that developed The Rose's theoretical model and related hedonic price analysis to supply, demand and market stability. Rabiega et al. (Rabiega, Lin, & Robinson, 1984)have worked on the effect of a public housing project on present houses if built near them using hedonic price technique. In the case of regression analysis, having data that shows heteroscedasticity can ruin the results (at very least it can produce very biased coefficients). In fact, most of the early researches suffered from this problem which pushed Stevenson to re-examine heteroscedasticity in hedonic house price models and try to eliminate it and in fact, his approach was successful in reducing heteroscedasticity (Stevenson, 2004). During the same year Bin (Bin, 2004) introduced a newer concept, a hybrid regression model called semi-parametric (non-linear) regression that can estimate a hedonic price function. To evaluate this semi-parametric function, Bin compared the performance of this model to the traditional parametric ones and the result was in favour of his model that performed well in both in-sample and out-of-sample predictions. He was also the first researcher to incorporate location data from geographic information system (GIS) and incorporate them into his regression model. More research was done by Kim and Park (Kim and Park, 2005) That identified the spatial pattern of housing price changes and their determinants in Seoul and its neighbouring new towns. Regression models were improved over time and the most significant improvement was done by Osland that applied semi-parametric analysis, Geographically-weighted regression (different locations have different weights in the equations) and spatial econometrics, all together in the process of hedonic house price modelling (Osland, 2010). Osland was successful in giving more weight in his analysis to the economically important areas through a comprehensive analysis, a breakthrough and a solution to a problem that many researchers faced before him. However, and as mentioned at the beginning of this proposal, regression models face many challenges and suffer from many limitations related to assumptions and estimation. These challenges are the identification of supply and demand, market instability, choice of independent variables, choice of hedonic function and market segmentation (Park and Kwon Bae, 2015). Hence the need for a "new" method than can predict the price of a property more efficiently and with less problems and effort, a new method that allows researchers to analyse and understand the relations between variables even if the relations were highly non-linear (Kuşan, Aytekin and Özdemir, 2010).

2.2 Machine Learning

2.2.1 Background

Compared with regression methods, machine learning methods are more suited for capturing relations between space and time and that is due to their non-assumable and much adaptive properties which allow them to capture the highly non-linear relations. The two most popular approaches for prices predictions are ANN (Artificial Neural Network) and SVM (Support Vector Machine) (Cheng et al., 2014). Machine learning was being used in many sectors like business, engineering, Physics and statistics to acquire knowledge and predict the future and recent studies have used machine learning techniques to predict housing prices instead of regression.

2.2.2 SVM

Using the statistical learning theory, SVMs were created to solve the problem of the estimation of multi-dimensional function (Vapnik, 2000). SVMs can also be grouped into two main categories: Support Vector Classification (SVC) and Support Vector Regression (SVR). While the first one is used to address classification problems (Abedi, Norouzi and Bahroudi, 2012), the last one was made to handle approximations of functions (Smola and Schölkopf, 2004). SVMs have the ability to avoid local optimal solutions and they can perform very well with space-time series data, contamination prediction, weather predictions and forest fire predictions (Lu and Wang, 2005; Pozdnoukhov et al., 2011). One downfall for SVMs is that they are unable to explain the relationships between different variables and it is hard to know the parameters to be used like type of kernel or slack variable. An early research done by (Lam, Yu and Lam, 2009) showed that SVM, with radial basis kernel function (RBF), can be used to predict market prices with results solid enough to consider SVM as a reliable tool to predict market prices (The parameters in this model were optimised to find the best results). Hence our choice of SVM.

2.2.3 ANN

ANN tries to simulate the functioning of biological nerves (neurons) in the human body and find a mathematical equation of information processing (Kohonen, 1988). Depending on their internal structure, ANNs are divided into: feedforward neural networks (Moody and Darken, 1989), feedback neural networks (Elman, 1990) and self-organising ones (Kohonen, 1982). The first two are always used for predictions analysis while the last one is used for cluster analysis. ANNs are dealt with as "black boxes" so it will be difficult to explain the domain knowledge.

2.2.4 Related Work

In this section we will choose the most related papers to our research, present the work done and build on it hoping that we will achieve best results with our research.

SVM: we will start with SVM as it is one of the two algorithms we chose to forecast the price of a property, after providing it with structural attributes like number of rooms, bathrooms etc. and distance to the nearest school. We hope for better results in our research as we are adding distance to nearby school, an attribute we believe people seriously consider before buying a house. As mentioned in the last section (Lam, Yu and Lam, 2009) proved SVM could a powerful tool that even outperformed ANN on the set of data they used. A newer research that is related to our study is the work done by (Oladunni and Sharma, 2017) where they used SVM-R to forecast the housing prices in the United States of America. Locational attributes were not present in this study either which presented us with an opportunity to include such a feature in our study and see how our algorithm will do. They also compared SVM with another machine learning algorithm called KNN and the results were in favour of SVM scoring 87% for performance while KNN scored 83%.

ANN: ANN is the second algorithm we will be using in our forecasting model. ANN recently gained a lot of attention in research for the positive results they are producing. (Kauko, Hooimeijer and Hakfoort, 2002) tested applying neural network with the housing market in Finland. Their results showed that different dimensions of housing sub-market formation could be uncovered by revealing the relations and patterns hidden in the datasets. Furthermore, they showed the capability of two neural network algorithms: self-organizing map and learning vector quantization. Khalafallah used ANN to predict the housing prices in the United States of America. The ANN model he used was a feed-forward back-propagation multilayer perception networks using NeuroSolutionsis. The reason behind his choice

was because this type of neural nets is believed to be very solid in estimating almost any input/output (Khalafallah, 2008). The results khalafallah got were very encouraging with a model testing showing some error between -2% and +2%. He used a different approach than the one we are going to follow as our research considers proximity to schools an important factor that should be included in the study. Instead he used variables like median house price compared to previous year (MdCh), average days a house spends on the market, the volume of inventory (Inv) and no spatial attributes.
A more recent work, done by (Chiarazzo et al., 2014) in the urban area of Taranto (Italy), where a number of dummy variables were used as independent variables, provided a solid proof that ANN can be used for this kind of forecasting. They had a large number of attributes that included structural variables, number of people in a zone, measurements of polluters found in air, number of employed in a zone and an attribute that we are mostly interested in which is distance to industrial center in km from the property. At the end of the training procedure, for each input sample, an output (estimated house price) has been estimated by the ANN model. Then, the obtained results have been compared with the target values (actual house price). The correlation coefficient R was close to 1 with the training set while R was 0.83 for the testing set which showed a good fit. The most significant variables appear to be those related to property special features such as proximity to the beach or presence of a garden or terrace which is a good indication that our choice of proximity to schools as an independent variable will have a significant positive impact on the performance of our model.

2.3 Critical Analysis of Existing methods and Niche Research

The analysis of what preceded can be summarized as follows:

- Regression and statistical models are capable of interpreting and predicting statistical inferences only if they had solid hypotheses, that is premises and assumptions must be correct. In fact, when used with space-time cases where non-linear relationships must be represented, like when predicting the housing prices, statistical methods proved to have low reliability. Domain knowledge and strong background in statistical modelling are required which present challenges for researchers when dealing with such methods.

- Machine learning methods have the power to estimate any non-linear relationship, even if there was only a little domain knowledge. Despite the fact they don't offer great interpret-ability, they are recently widely researched.

- The methods of benchmarking and gauging the performance of different algorithms are different and specific to each paper which includes some uncertainty to the concept of comparing these algorithms. It would be interesting to see if different research papers used the same benchmarking techniques.

- SVM and ANN are one the most popular algorithms used for forecasting housing prices therefore we will use both algorithms in our study and adopt the best performing one as a tool for future predictions. forecasting.

- Distance to schools will be included in our model as previous research proved the strong contributions of these special features to the models used (Niche Research).

In the next chapter we will discuss our research method and specification including all steps taken to acquire the data through datasets and APIs, clean it and use it to build our two models (SVM and ANN). Also we will introduce evaluation methods that will allow us to judge our work and results. Finally a Gantt chart will be added at the end to show the time allocated for the tasks to be carried out during the project.

Chapter 3

Research Methods and Specifications

3.1 CRISP-DM

CRISP-DM method is going to be followed throughout our study. It is a process model developed in 2000 by CRISP-DM consortium (Chapman et al., 2000) that describes the approaches we are going to use to tackle all the issues we might face during our project. Its process outlines all steps involved in performing data science activities from business need to deployment and most importantly it shows how repetitive this process is and that we never get thins perfectly right. Figure 3.1 (Chapman et al., 2000) summarizes the process that our project will go through from start to finish. It is an overview of the life cycle of a our project. It includes the phases of our work, its respective tasks, and the relationships between these tasks.

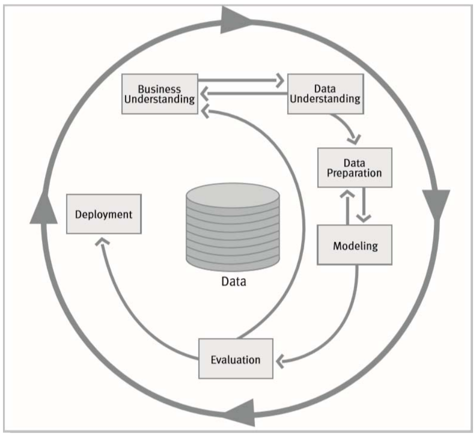


FIGURE 3.1: CRISP-DM (Chapman et al., 2000)

- Business understanding: This is the initial phase that focuses on understanding our work objectives and needs from a business point of view and then convert this knowledge into a machine learning problem definition and an initial plan.This phase has been addressed comprehensively during the first submission and it is closely related to the next step which is the data preparation, as in order for us to understand data properly we need to understand the business needs first.

- Data understanding: R will be used to get familiar with the data and to check if there are obvious problems like missing values, strange characters that need to be removed or altered. This process will also give us the chance to see if there are obvious insights into the data or to capture subsets that could lead to important hypotheses for hidden information.
- Data preparation: This is usually the longest part of any project. It is a very important step that we need to explain it thoroughly in the next section (design). This phase starts with downloading the data from their sources using R and then cleaning it from missing values and weird characters. Finally the clean data will be aggregated to be used in the model.
- Modelling: Once we have our data ready and cleaned then it is time to start building our model. For reasons explained before It is already decided to use ANN and SVM as modelling techniques. In this phase we will apply these two models and we will tweak their parameters to find the optimal performance. SVM can be tweaked using different Kernels and ANN will be tweaked using different combinations of layers.
- Evaluation: The nature of our research question requires having solid evaluation techniques that allow us to gauge the performance of our models. At this stage the model obtained is thoroughly evaluated and the results are checked to confirm they fulfil the business requirements.
- Deployment: Creation of the model is usually not the end of the project. The knowledge acquired needs to be presented in way so other researchers can use it and build on it for future research.

3.2 Design

In this section we will provide a more thorough explanation and design documentation of the major design decisions.

3.2.1 Data Preparation

This is a very important step in the life cycle of our project. In any machine learning project, data preparation is considered the longest and the most important steps to go through as the performance of the model relies heavily on the quality of data fed to it, which means if our input was clean then we will have good reliable results. Daft.ie keeps records of all properties sold in Ireland between 2011 and 2017 (around 18,000 records per year) that includes the selling prices along with structural attributes like the type of the property, number of bedrooms, number of bathrooms, availability of garden and surface area of the property and whether it is a first or second hand. Google maps will also be used in our work to retrieve coordinates of all properties included. Also an additional dataset downloaded from data.gov.ie and that has the coordinates of each school in Dublin city will be added to the rest of the datasets to be analysed. The three sources of data will be targeted by R to download the data and do all the cleaning needed. Also R will be used to calculate the distance between the property sold and the nearest school and create an additional column called distance to school. At the end of this process all clean datasets will be joined to from only one.

Figure 3.2 below shows the flow of data from the sources to R.

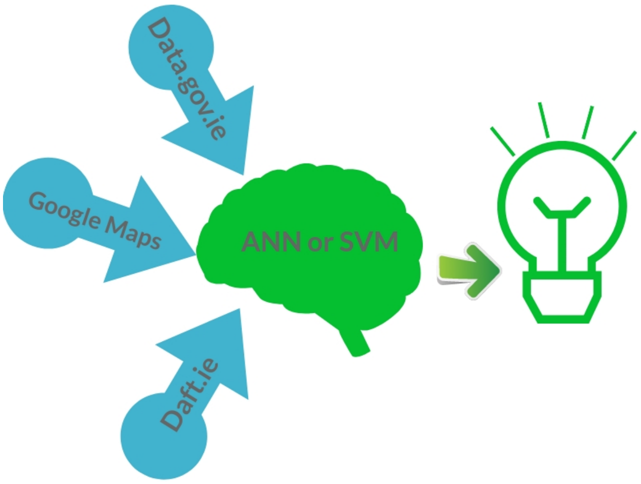


FIGURE 3.2: Data Flow

Once our data is downloaded and cleaned, the three different datasets will be joined to form only one. The new dataset now has the structural attributes mentioned above in addition to the coordinates of schools across Dublin and coordinates of each housing property sold in the city. The fact we have two sets of coordinates on the same dataset allows us to calculate the distance between the property and the surrounding schools and then let R sort the distances and only keep the smallest one. The distance feature now created will be added to the dataset as another variable that will be used in our algorithm. Also, normalization test will be performed using Chi-Square or Q-Q plot to check whether our data is normalised and in case it wasn't, R can normalize it easily. Also, we need to check if our data is imbalanced, which could bias our results and our algorithm will loose on performance.

3.2.2 Modelling

ANN and SVM are the two modelling algorithms we are going to be using for our work for reasons discussed previously in this paper. We will be using R throughout our modelling process that offers two packages for these algorithms and they are called (neuralnet) for ANN and (e1071) for SVM.

With ANN we are going to adopt feed-forward back-propagation neural networks as they proved to be very robust in predicting the outputs from any input (Khalafallah, 2008). (Chiarazzo et al., 2014) also confirmed that this kind of structure for ANNs is reliable when predicting prices of properties. We will also include a training function that updates the values of weights and biases using Levenberg-Maruquardt optimization. The rest of the parameters like the number of hidden layers and nodes will be tweak according to the work done by (Lin and Chou, 2016) in order to get the best performance, where they developed an equation that will provide the best choice of parameters to be used in a neural network but also we will change parameters manually and repetitively to try and improve performance. Data will be normalised as previously mentioned, which is a very important step when dealing with neural network and a very complex training task. Data will be divided into training and testing sets and lastly the neural networks will be fit to produce our model.

On the other hand, when dealing with SVM there is no set rule what kind of parameters would give the best results. Package (e1071) offers many kernels that we can change until we get the best performance. Also we will check whether dimensionality reduction can help our model. We have numeric data so that suggests PCA. Running PCA on the whole dataset could bias our performance measure that is why we should make a stratified holdout with a ratio 80:20 (training to test). Vector loadings will be gotten after this steps and then used in the model to train it. We can also try improving the performance of our model by changing the number of PCAs that will be included. It is important to spend as much time need here in this process as possible and check the trade off of performance vs. number of PCs vs. time needed to build and evaluate the model. The evaluation of the performance will be highlighted in the following section

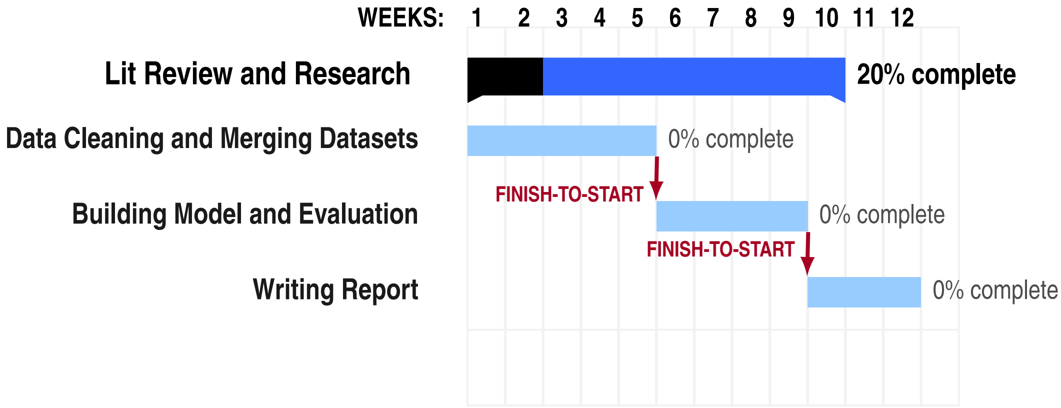
3.2.3 Evaluation

Because we are trying to prove that adding distance to schools will improve our model, we need to compare each of the two models (ANN and SVM) with and without the distance. As far as the technical side of the evaluation method there are many evaluation methodologies available and we will choose the following ones:

ANN: A good way to evaluate the performance of an ANN algorithm is by using the MSE or the mean squared error which is a measure of how much our predictions are far away from the real data. Also we include a linear model which acts as a control that allows us to compare the MSEs between our ANN model and a linear one. For MSE, the less the better. Also to be sure of our results we will use another method of evaluation called cross validation. It works by splitting the data between training and testing set, fitting the model, testing the model on the testing test and then calculate the prediction error. The process is repeated k times then an average of all errors will be calculated to get a grasp about how the model is doing.

SVM: It offers a function called performancesvm(model, test) which you can run on any svm model you build and it will return a percentage which is the accuracy of your model. Here the higher the number the better. Here we have to be careful as it is very important to take the data used into consideration and not depend on accuracy as the sole gauge of how our model is doing because an imbalanced data could produce a false high accuracy. Therefore, specificity and sensitivity could be check as well to have a better gasp of the real performance.

3.3 Gantt's Chart



The Gantt's chart shown above is a series of horizontal lines made to show the progress of work done on my thesis in relation to the planned work over 12 week period.

At the top of the chart we can see the "Lit Review and Research" bar that shows 20% completion and that is because the work done on this proposal is considered part of the literature review that needs to be done for my thesis. Also we can see that the literature review and the research will start on the first week and continue until week 10, which could be explained by the fact we might need to read more papers about some challenges we might face when building our model hence the time allocated for research is around 10 weeks.

Data cleaning and preparation will take around four weeks to be done as the size of the datasets that need to be downloaded is quite large with a lot of attributes and in separate sets and we will be working through APIs from daft.ie and Google maps which might lead to a lot of missing or non conformed values. Also the coordinates that need to be downloaded from Google maps need to matched with the coordinates downloaded from data.gov.ie in R.

The process of building a model is a process that relies heavily on the way our data was cleaned. R offers many packages that allow us to use any model we want with the data. This has made building models an easier process but keeping in mind that the important task of finding the right parameters will

have to be done manually to achieve the best performance and that could be a lengthy process, hence the 5 week allocated time for this phase.
The last process is writing the report. This process is scheduled to take around three weeks to be completed. Throughout our research, notes will be taken of all steps, procedures and results. Our report will include part of the literature review done on this proposal, detailed reviews of any additional related work we will read about and a thorough explanation of procedures and setps we followed.

3.4 References

Abedi, M., Norouzi, G. H. and Bahroudi, A. (2012) 'Support vector machine for multi-classification of mineral prospectivity areas', Computers and Geosciences. doi: 10.1016/j.cageo.2011.12.014.

Bin, O. (2004) 'A prediction comparison of housing sales prices by parametric versus semi-parametric regressions', Journal of Housing Economics. doi: 10.1016/j.jhe.2004.01.001.

Brown, J. N. and Rosen, H. S. (1982) 'On the Estimation of Structural Hedonic Price Models', Econometrica, p. 765. doi: 10.2307/1912614.

Chapman, P. et al. (2000) 'Crisp-Dm 1.0', CRISP-DM Consortium, p. 76. doi: 10.1109/ICETET.2008.239.

Cheng, T. et al. (2014) Spatiotemporal data mining, Handbook of Regional Science. doi: 10.1007/978-3-642-23430-968.

Chiarazzo, V. et al. (2014) 'A neural network based model for real estate price estimation considering environmental quality of property location', Transportation Research Procedia. Elsevier B.V., 3(July), pp. 810–817. doi: 10.1016/j.trpro.2014.10.067.

Elman, J. L. (1990) 'Finding structure in time', Cognitive Science. doi: 10.1016/0364-0213(90)90002-E.

Gu, J., Zhu, M. and Jiang, L. (2011) 'Housing price forecasting based on genetic algorithm and support vector machine', Expert Systems with Applications. Elsevier Ltd, 38(4), pp. 3383–3386. doi: 10.1016/j.eswa.2010.08.123.

Kauko, T., Hooimeijer, P. and Hakfoort, J. (2002) 'Capturing housing market segmentation: An alternative approach based on neural network modelling', Housing Studies. doi: 10.1080/02673030215999.

Khalafallah, A. (2008) 'Neural Network Based Model for Predicting Housing Market Performance', 13(October), pp. 325–328.

Kim, K. and Park, J. (2005) ‘Segmentation of the housing market and its determinants: Seoul and its neighbouring new towns in Korea’, Australian Geographer. doi: 10.1080/00049180500150019.

Kohonen, T. (1982) ‘Self-organized formation of topologically correct feature maps’, Biological Cybernetics. doi: 10.1007/BF00337288.

Kohonen, T. (1988) ‘An introduction to neural computing’, Neural Networks, 1(1), pp. 3–16. doi: 10.1016/0893-6080(88)90020-2.

Kuşan, H., Aytekin, O. and Özdemir, I. (2010) ‘The use of fuzzy logic in predicting house selling price’, Expert Systems with Applications, 37(3), pp. 1808–1813. doi: 10.1016/j.eswa.2009.07.031.

Lam, K. C., Yu, C. Y. and Lam, C. K. (2009) ‘Support vector machine and entropy based decision support system for property valuation’, Journal of Property Research, 26(3), pp. 213–233. doi: 10.1080/09599911003669674.

Lin, T.-H. and Chou, J.-H. (2016) ‘Study on the optimal parameters of artificial neural networks by applying uniform design’, 2016 International Conference on System Science and Engineering (ICSSE), (1), pp. 1–2. doi: 10.1109/ICSSE.2016.7551640.

Lu, W.-Z. and Wang, W.-J. (2005) ‘Potential assessment of the “support vector machine” method in forecasting ambient air pollutant trends’, Chemosphere. doi: 10.1016/j.chemosphere.2004.10.032.

Moody, J. and Darken, C. J. (1989) ‘Fast Learning in Networks of Locally-Tuned Processing Units’, Neural Computation. doi: 10.1162/neco.1989.1.2.281.

Mukhlishin, M. F., Saputra, R. and Wibowo, A. (2017) ‘Predicting House Sale Price Using Fuzzy Logic , Artificial Neural Network and K-Nearest Neighbor’, (1), pp. 171–176.

Oladunni, T. and Sharma, S. (2017) ‘Hedonic housing theory - A machine learning investigation’, Proceedings - 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016, pp. 522–527. doi: 10.1109/ICMLA.2016.103.

Osland, L. (2010) ‘An Application of Spatial Econometrics in Relation to Hedonic House Price Modeling’, Journal of Real Estate Research, 32(3), pp. 289–320. doi: 10.5555/rees.32.3.d4713v80614728x1.

Park, B. and Kwon Bae, J. (2015) ‘Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data’, Expert Systems with Applications. Elsevier Ltd, 42(6), pp. 2928–2934. doi: 10.1016/j.eswa.2014.11.040.

Pozdnoukhov, A. et al. (2011) ‘Spatio-temporal avalanche forecasting with Support Vector Machines’, Natural Hazards and Earth System Science. doi: 10.5194/nhess-11-367-2011.

Smola, A. J. and Schölkopf, B. (2004) ‘A tutorial on support vector regression’, Statistics and Computing. doi: 10.1023/B:STCO.0000035301.49549.88.

Stevenson, S. (2004) ‘New empirical evidence on heteroscedasticity in hedonic housing models’, Journal of Housing Economics. doi: 10.1016/j.jhe.2004.04.004.

Vapnik, V. N. (2000) The Nature of Statistical Learning Theory, Springer. doi: 10.1109/TNN.1997.641482.