University of Warsaw
Faculty of Economic Sciences

Working Papers
No. 5/2021 (353)

THE APPLICATION OF MACHINE LEARNING ALGORITHMS FOR SPATIAL ANALYSIS: PREDICTING OF REAL ESTATE PRICES IN WARSAW

Dawid Siwicki

Warsaw 2021

WORKING PAPERS 5/2021 (353)

The Application of Machine Learning Algorithms for Spatial Analysis: Predicting of Real Estate Prices in Warsaw

Dawid Siwicki

University of Warsaw, Faculty of Economic Sciences

Corresponding authors: dawid@siwicki.co

Abstract: The principal aim of this paper is to investigate the potential of machine learning algorithms in context of predicting housing prices. The most important issue in modelling spatial data is to consider spatial heterogeneity that can bias obtained results when is not taken into consideration. The purpose of this research is to compare prediction power of such methods: linear regression, artificial neural network, random forest, extreme gradient boosting and spatial error model. The evaluation was conducted using train, validation, test and k-Fold Cross-Validation methods. We also examined the ability of the above models to identify spatial dependencies, by calculating Moran's I for residuals obtained on in-sample and out-of-sample data.

Keywords: spatial analysis, machine learning, housing market, random forest, gradient boosting

JEL codes: C31, C45, C52, C53, C55, R31

INTRODUCTION

Hedonic price models are widely considered as a starting point in researches related to real estate analysis. However, investigation of impact only of the indoor attributes of the property on its price can lead to wrong conclusions. We can assume that prices are affected by neighbourhood amenities and that assumption allows to consider real estate market as spatially autocorrelated. In order to investigate the spatial effects, researchers proposed spatial models that explain the determinants of property prices including spatial autocorrelation.

Although, spatial models allow to investigate spatial dependencies there are also some limitations that should be considered. Firstly, it is unnecessary normality assumption and misspecified linear form of relationship as limitations of SAR models (Pinkse & Slade, 2009). Secondly, McMillen (2010) noticed also unreasonable assumption that the appropriate model structure is known beforehand. Thus, Hong et al. (2018) proposed machine learning methods, specifically neural networks, as an alternative for real estate prices modelling. Hedonic models simplify complexity and non-linearity of real-world data. Also, Selim (2009) found out potential non-linearity of the hedonic functions and proposed artificial neural network as an alternative solution. Advanced machine learning algorithms do not need prior assumptions on functional form and learn from the provided data to solve more complex problems. Additionally, necessity of including spatial weights matrix, when applying spatial models, causes problems with time and space complexity. McMillen (2010) pointed out that maximum likelihood method that is widely used in spatial models performs well on large data and researches related to spatial models are usually based on small-to-medium datasets.

In this paper we attempt to investigate the potential of machine learning algorithms in comparison to spatial models in terms of the prediction performance. We consider linear regression, artificial neural network, random forest and extreme gradient boosting as possible alternatives for price prediction on real estate market in Warsaw, Poland in 2018. **The hypothesis examined in this paper is whether it is possible to specify algorithm that outperform spatial autoregressive models in prediction of the real estate prices.** We also attempt to examine ability of the implemented methods to identify spatial correlations in provided data. **We assume that spatial techniques allow to avoid a bias caused by existence of spatial heterogeneity.**

Out of the multiple spatial models that can be deduced from Manski model and spatial weights matrices, Spatial Error Model with k-Nearest Neighbour Weight Matrix was chosen as a spatial model that perform best on provided data. The combination of *train, validation, test* and *k-Fold Cross-Validation* methods were applied to evaluate all introduced models.

The paper is structured as follows. First, we attempt to review the relevant research related to the spatial analysis of real estate prices and studies on application of machine learning methods to the purpose of the raised issue. Next, we present the process of obtaining and pre-processing data. Then, in the last chapter we introduce and apply the algorithms and investigate its ability as a prediction tools for real estate market.

1. LITERATURE REVIEW

Approaches for real estate market analysis

The studies related to real estate market analysis consider multiple number of methods to understand how housing prices can be explained. It was proposed to apply hedonic price models for real estate market (Rosen, 1974) and that method was widely used until 1990's. Hedonic price model is the approach that attempt to define price of a specified good as a function of characteristics of this good. However, traditional regression estimations may be biased when modelling "of geographically referenced data due to spatial effects, namely spatial dependence and spatial heterogeneity" (Can, 1992). Influence of the location should be included in process of estimation of property prices. It can be assumed that neighbourhood is developed in the same time, so structural characteristics are similar, and amenities are shared by adjacent properties (Basu & Thibodeau, 1998). That leads to existence of spatial heterogeneity that is not explained properly by simple regression methods. Avoiding the spatial effects causes the model to tend to overestimate the impact of structural and neighbourhood attributes (Yu, Wei & Wu, 2007). Out of the methods that can be derived from Manski model and were classified by Elhorst (2010), spatial lag and spatial error models are the most commonly used in relevant research. Nevertheless, simple regression model using least squares method is widely used as a reference model for spatial approaches in research that investigate real estate market. This makes it possible to conclude that in general spatial models perform better than OLS also in terms of predictive accuracy (Yu, Wei & Wu, 2007). The studies that investigate potential of spatial lag and spatial error models in comparison to traditional hedonic model that uses least squares, confirm that both proposed spatial methods outperform simple regression model (Conway et al., 2008; Osland, 2010; Wilhelmsson, 2002; Zhang et al., 2015). Recently, the potential of advanced machine learning methods in real estate market analysis is widely taken into account. It was confirmed that methods that can be deduced from Manski models are efficient in eliminating spatial autocorrelation. On the other hand, these models are similar to simple regression in assumption on linearity of the relationship between provided data. This

assumption is naive when handling real-world scale problems and leads to obtaining biased results. Recent studies propose a number of machine learning algorithms that can handle more complex problems and do not need any prior assumptions on functional form of the model. Neural network is the method that can be a powerful alternative for estimation property prices based on its attributes (Limsombunchai, 2004; Selim, 2009) and additionally can investigate a spatial correlation based on a satellite images of property neighbourhood (Bency et al., 2017). Also, tree-based methods can be considered as an alternative for spatial models. Random Forest that is ensemble machine learning algorithm, were examined as a predictive tool for spatial structured problems. It was confirmed that this method can outperform both, linear regression (Hong, Choi & Kim, 2019) and spatial models, namely Kelejian-Prucha (SAC) model (Song et al., 2017). However, SAC model showed ability for explaining spatial structure of the provided data. Additionally, gradient tree boosting algorithm can outperform other methods in terms of predictive power (Peng, Huang & Han, 2019).

The determinants of real estate prices

Features that are widely considered as factors that determine real estate prices, can be divided into three main categories (Espinoza, 2019). Firstly, structural attributes that describe indoor specification of a property and usually can be found in sale offer. The widely used features in related studies are *living area* and *age* of a property that are considered in most of mentioned research (Conway et al., 2008; Limsombunchai, 2004; Wilhelmsson, 2002). Despite of potential correlation with living area, number of rooms in a property is frequently taken into account in studies that investigate determinants of housing prices (Bency et al., 2017; Selim, 2009). Secondly, neighbourhood attributes describe amenities that can be found in property neighbourhood. A number of studies investigate influence of access to a garage (Basu & Thibodeau, 1998; Limsombunchai, 2004; Osland, 2010) or a parking lot (Hong, Choi & Kim, 2019) on real estate prices. Also, number of specified amenities in specified distance to property can be considered as neighbourhood features. Finally, locational attributes that attempt to show location of a property. Geographical coordinates (latitude and longitude) are included in studies that investigate potential of machine learning algorithms (Bency et al., 2017; Hong, Choi & Kim, 2019). Usually researchers attempt to present location by considering distances to nearest amenities. Widely used are distances to nearest subway station (Espinoza, 2019; Hong, Choi & Kim, 2019) and nearest green park (Conway et al., 2008; Limsombunchai, 2004). The impact of distance to nearest schools and hospitals are also examined in studies related to real estate market analysis (Hong, Choi & Kim, 2019; Sun, Wang & Li, 2016). Due to the potential noises

and negative influence on life comfort, distance to the airport was examined in some research (Cohen & Coughlin, 2008).

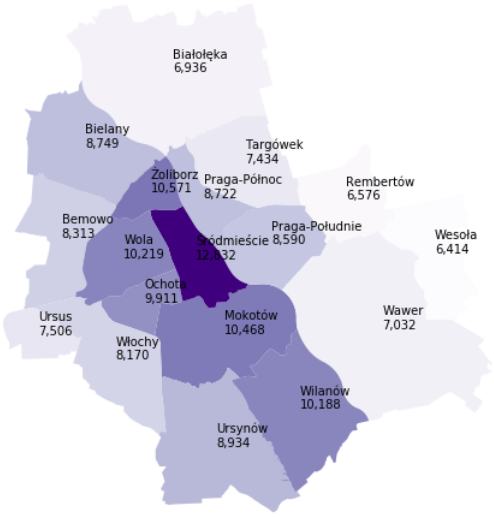
2. RESIDENTIAL PROPERTY MARKET IN WARSAW

For the purposes of the following research, the data was collected from one of the most popular webpages with advertisements of real estates and plots for rent and sale in Poland – www.morizon.pl. Data about sales offers on secondary residential property market in Warsaw was gathered on November 2018 using web scraping methods in Python. Additionally, Google Places Nearby Search API was used to gather the data about Points of Interests (POIs) in Warsaw including: theatres, medical clinics, grocery stores, restaurants, parks, Park&Ride parkings, night clubs, malls, bars, subway and railway stations, bus and tram stops and Warsaw Chopin Airport (WAW). Research includes also data shared by City of Warsaw and its institutions – educational establishments and Veturilo - public bicycle sharing system in Warsaw - stations. Geographical coordinates of real estates were obtained based on gathered addresses and using Google Geocoding API.

The essential in context of big datasets is preliminary data analysis. Due to the fact that scraped data takes raw form, at the beginning data was cleaned and valuable information was obtained from unstructured text variables. After cleaning and handling with missing data by different imputation and deletion methods, feature engineering process was applied to create distance-based variables basing on localizations of Points of Interest. Moreover, one-hot encoding was applied where it was needed for quality variables. After data pre-processing, final dataset contains 21 226 observations and 104 variables, including target variable and estates' latitude and longitude that were used in calculations of distances from estates to POIs and computing spatial wages matrices for purposes of spatial models.

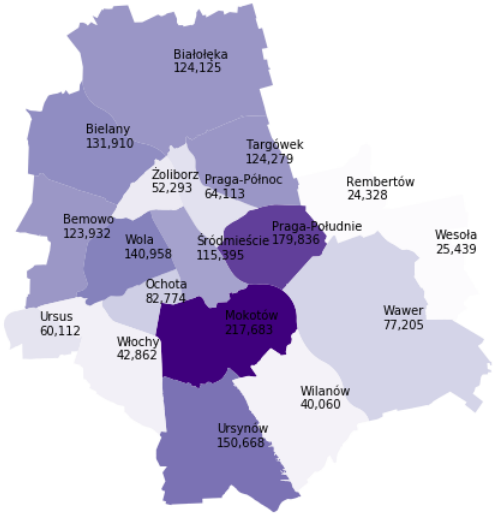
Warsaw, as a capital and the largest city in Poland, has the most developed real estate market in the country. According to the data shared by City of Warsaw, there were 1 777 973 inhabitants at the end of 2018, including negative birth rate (- 1 603) and positive migration balance (9 730). The number of 23 430 new flats had been put into service and average price per square meter in comparison to 2017, increased to amount of PLN 9 543.

Figure 1 The average prices of square meter in Warsaw's districts (2018).



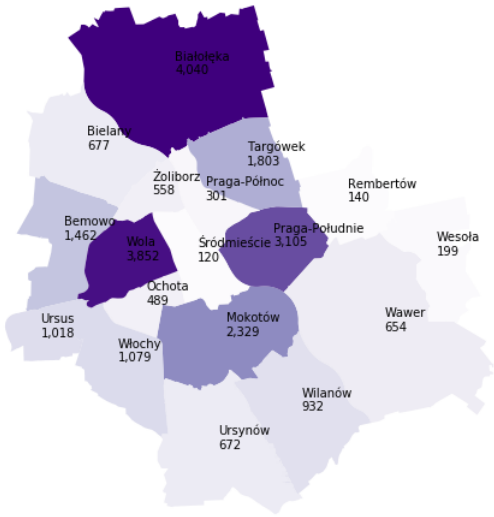
Source: Own preparation.

Figure 2 Number of inhabitants in Warsaw's districts (2018).



Source: Own preparation.

Figure 3 Number of new flats put into service in Warsaw's districts in 2018.



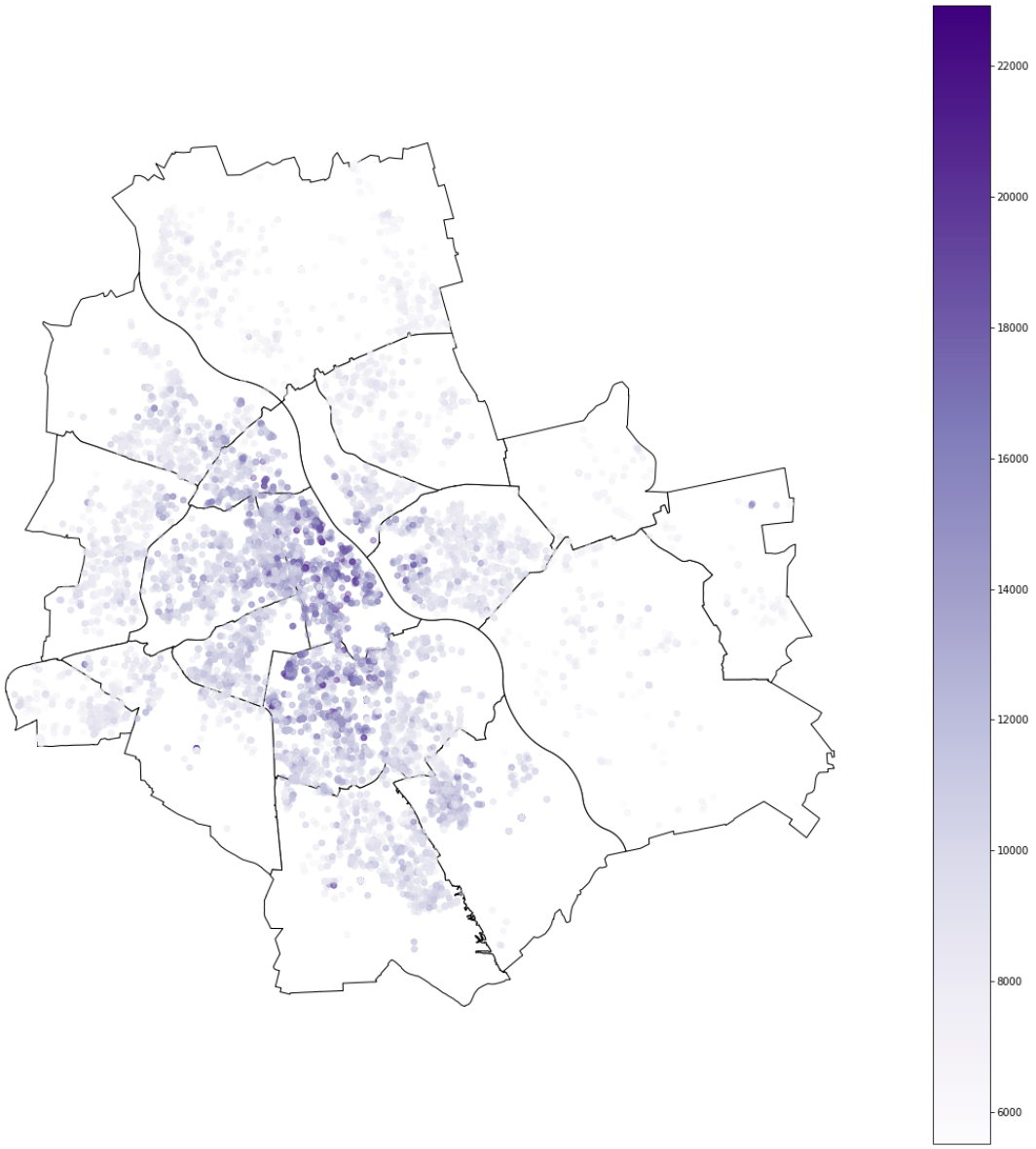
Source: Own preparation.

According to EY's report - The Polish Real Estate Guide 2018 Edition - immigration, the highest level of income and the lowest unemployment rate in Poland are the main factors influencing on demand on the market. Increasing number of office and service centers investments attract people, both from abroad and other regions of Poland. Increasing demand

for residential real estates results in rise of the prices (about 8,4% year to year in 2018) – the most expensive and prestigious offers exceed the amount of PLN 15 000 per square meter.

In this paper, the **dependent variable is offer price per square meter** (PRICE_M2) expressed in Polish Złoty (PLN) and was obtained by dividing total offer price by property area expressed in square meters (AREA). Estimation of total price can be biased by strong correlation with area – one can assume, the larger estate, the higher price. In order to focus in this study on impact of localization and available amenities, price per square meter was approximated by the machine learning models. In used dataset, the dependent variable is distributed in range between 5509 and 22 929 PLN, with mean equal to 9924 PLN and median equal to 9399 PLN. One can notice, average price per square meter in used dataset is close to average price shared by City of Warsaw.

Figure 4 The spatial distribution of real estates and prices per square meter in obtained data.



Source: Own preparation.

Hedonic analysis makes use of structural attributes of real estates and in this paper impact of both quality and quantity **property characteristics** were considered. Among quality features were included such variables as: CONDITION (binary, equal to 1 if estate need renovation or is in builders finish, 0 in other cases), CELLAR (binary, equal to 1 if cellar is included) and GARAGE (dummy, equal to 1 if parking place in garage is included, 2 if outdoor parking place is included, 3 if more than one parking place are included, 0 if parking place is not included). As quantity features were included: AREA (real estate area in square meters), AGE (age of real estate in years), FLOOR (number of floor, on which real estate is based),

FLOORS_CNT (number of floors in the building in which real estate is based) and ROOMS_CNT (number of rooms, which real estate consists).

Also, **neighbourhood and accessibility** data were included. Based on geographical coordinates of real estates and Points of Interest, distance-based features were calculated. As it was mentioned, neighbourhood of such types POIs was considered: theatres, medical clinics, grocery stores, restaurants, green parks, Park&Ride parkings, night clubs, malls, bars, subway and railway stations, bus and tram stops, airport, educational establishments and public bicycle sharing system stations. For nearly all included types of POIs, five variables were approximated: distance to the nearest POI of given type (e.g. MALL_DIST) expressed in meters, number of given type's POIs in range of 500m (e.g. MALL_500M_CNT),, number of given type's POIs in range of 1000m (e.g. MALL_1000M_CNT), number of given type's POIs in range of 3000m (e.g. MALL_3000M_CNT), and number of given type's POIs in range of 5000m (e.g. MALL_5000M_CNT). In case of Vistula river and Warsaw Chopin Airport only distance to the POI was calculated (VIST_DIST and AIRPORT_DIST).

3. METHODOLOGY

Below we presented methods applied in the following paper including their assumptions, specificity, hyperparameters tuning process and models assessment. Analysis starts with simple linear regression, which is widely used in researches that concern spatial relationships on real estate markets. Then introduces machine learning algorithms like Random Forest, Artificial Neural Network and Gradient Boosting, and finally Spatial Error Model, which represents spatial modelling methods and considers spatial effects.

Linear Regression

Linear models are one of the most important tools in statistics and although they had been developed in pre computer age, nowadays are still widely used. In researches associated with spatial analysis, linear regression is commonly used in comparison to spatial models (Conway et al., 2008; Osland, 2010; Wilhelmsson, 2002; Zhang et al., 2015).

The linear regression model is formulated as follows:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \qquad (1)$$

where β_j are unknown coefficients and X_j are independent variables.

The model is linear in the parameters and there is assumption that regression function E(Y|X) is linear or that the linear model is a reasonable approximation (Hastie et. el. 2009). Based on training set of data parameters are estimated. In the following research it is applied the most popular method of estimation – Least Square, in which coefficients are picked to minimize the residual sum of squares:

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 \quad (2)$$

For the purposes of the following research, forward feature selection is applied for linear regression to choose best combination of variables. Final linear regression is estimated using 26 most powerful variables chosen among 102 features in original dataset.

Artificial Neural Network

The idea of artificial neural networks (ANNs) is inspired by the way how human brain processes information. One can think, neural network is a set of interconnected neurons that make series of transformations on the input and produce its own understanding of it as an output. As a black-box algorithm, Neural Networks do not need any prior assumptions about functional relationship between the input and the output. Due to that fact and including multiple number of layers, ANN can solve not only problems where relationship is linear, but also more complicated problems on high level of abstraction. Neural Network is most common machine learning algorithm that is used in studies related to real estate market analysis (Bency et al., 2017); Limsombunchai, 2004; Selim, 2009).

The class of Neural Network used in this paper is the feedforward network, also called the multilayer perceptron (MLP), which is perhaps the simplest form of neural network. The main assumption of MLP is that it consists at least three layers: the input layer and the output layer, and at least one hidden layer between them – by adding more layers and more units within a layer, a network can represent functions of increasing complexity (Bengio, Goodfellow & Courville, 2016). Starting with the input layer, each neuron is given a connection strength (called typically a weight) to each neuron in next layer and represents influence of one unit on another. The weighted sum is calculated including bias and result is passed forward to next layer. Moreover, each layer is given an activation function that is responsible for activation connected units when given threshold is exceeded. Process continues in each layer until output

vector is passed to the output layer. In order to minimize prediction error, backpropagation algorithm was used in network training. Algorithm proceeds backwards through the network and adjusts weights by calculation of the loss function gradient and reduction of error value.

Although neural networks do not need any assumptions on functional relationship between input and output, some assumptions about algorithm parameters have to be provided. Starting with arbitrary chosen parameters, most appropriate parameters are chosen during process called hyperparameters tuning. The very first challenge is to find most appropriate number of hidden layers and number of neurons in each layer – as it was mentioned, increasing number of layers influence on network complexity and allows solving more complicated problems. In general, number of layers and neurons cannot be calculated and can be specify using the trial and error method. After designing optimal architecture of the network, the best set of following hyperparameters were chosen: learning rate, optimizer, initializer, batch size and dropout rate.

The learning rate, also known as gamma, is a parameter responsible for speed of learning process and controls impact of loss function optimization on adjusting weights in backpropagation process. Low learning rate slows down convergence of the model, while too high value of gamma threatens passing global minimum. Optimizer is algorithm that optimizes loss function by adjusting weights. Activation function can be used in each layer, but one can consider efficiency and velocity. The weighted sum calculated in the layer is passed through the given activation function to determine the output of the layer. Batch size represents size of subsample used when adjusting weights in algorithm learning process to prevent overgeneralization of the model. Dropout rate contains information how many nodes are removed to avoid overfitting thus increasing generalization power. Too low value of dropout rate results in under-learning of the model.

Moreover, in case of neural network algorithm the good idea is to standardize inputs to have mean zero and standard deviation one (Hastie et al. 2009). It allows to treat inputs equally and choose meaningful range for starting weights. Such standardization is applied on input data for the purposes of the neural network training.

Random Forest

Random Forest algorithm is ensemble method based on decision trees and was proposed by Breiman in 2000. This method can be used for both classification and regression problems as well. Algorithm is constructed from multiple independent decision trees built on bootstrapped training dataset. Number of trees in the forest influences on prediction robustness and model

accuracy – the more trees, the higher accuracy and the more robust prediction is. The estimation of random forest is a result of aggregating trees' estimations by averaging.

The decision tree algorithm is constructed by recursive partitioning, which is step-by-step process of splitting each node – starting from the root node - into child nodes (leaves). In each node data is iteratively split on most powerful feature – in terms of model fit - until some stop rule is applied (Hastie et al. 2009).

The strength of a random forest algorithm is that this method handles missing data. The Breiman's procedure takes advantage of the proximity matrix, which measures proximity between pairs of observations to estimate missing values. The method produces highly accurate predictions and can handle a big number of variables without overfitting (Biau, 2012).

The potential of Random Forest in estimating housing prices were investigated in number of relevant studies (Bency et al., 2017; Hong, Choi & Kim, 2019; Song et al., 2017).

eXtreme Gradient Boosting

The concept of boosting is based on theoretical approach to converting weak learners into strong ones, where weak learners are defined as a classifiers at least slightly better than random guessing and strong learners are classifiers approaching to perfect classification (idea was originally introduced for classification problems and then extended to regression problems as well). In 1996, Shapire and Freund have been introduced first implementation of boosting - Adaptive Boosting (AdaBoost) algorithm, which is now considered as a special case of Gradient Boosting.

In the AdaBoost, first decision tree is trained on original data set where observations have equal weights. After decision tree training weights increase for observations which are difficult to classify and decrease for those that are easy to classify. Second model is trained on weighted data and the idea is to improve result from first decision tree. The idea is not to manipulate the base-learner to improve its performance but to manipulate the underlying training data by iteratively re-weighting the observations. Process is repeated for specified number of iterations and in each iteration new model computes the classification error from ensemble model based on decision trees from previous iterations. Result of final ensemble model is weighted sum of results of the previous trees.

AdaBoost is different to modern gradient boosting algorithm, mainly in terms of way of identification of shortcomings of decision trees. In case of Adaptive Boosting, shortcomings are identified by using high weight data points and gradients are used in the loss function to perform the same in case of modern gradient boosting.

eXtreme Gradient Boosting (XGB, XGBoost) is the implementation of gradient boosting algorithm included in the following research. The authors proposed two improvements to gradient boosting algorithm - a novel tree learning algorithm to handle with sparse data and weighted quantile sketch for approximate learning (Chen & Guestrin, 2016). Scalability of XGB is one of the most important advantages of this method and allows researchers to solve real-world scale problems. It was confirmed that gradient boosting can outperform other methods in estimation of housing prices (Peng, Huang & Han, 2019).

eXtreme Gradient Boosting is algorithm that hyperparameters tuning process has to be conducted to find best-performed model. Well-chosen parameters allow to handle with common problems in machine learning, such as overfitting or underfitting.

Spatial models

Modern classification of spatial models was presented by Elhorst in 2010. The starting point for his research was Manski Model, which includes three types of spatial interactions: an endogenous interaction, an exogenous interaction and spatial correlation effects. The Manski Model can be formulated as follows:

$$\begin{aligned} y = {} & \rho W y + \beta X + \theta W X + u \\ u = {} & \lambda W u + \varepsilon, \end{aligned} \tag{3}$$

where W is spatial weights matrix, ρ, θ, λ are parameters of spatial autocorrelation, Wy is spatial lag of dependent variable, WX are spatial lags of independent variables, u is model error, and Wu – spatial lag of model error.

Although, model offered by Manski is good starting point to spatial modelling, it is not widely used because of over-specification. Imposing zero-restrictions on Manski model, reduced spatial models can be deduced: Spatial Durbin Error Model (SDEM, ρ = 0), Kelejian-Prucha Model (SAC, θ = 0) and Spatial Durbin Model (SDM, λ = 0). Before Elhorst's research on spatial models classification, researchers had been widely using three models on lower level of complexity: Spatial AutoRegressive Model (SAR, θ = λ = 0), Spatial Error Model (SEM, θ = ρ = 0) and Spatial Lag of X Model (SLX, λ = ρ = 0).

In contrast to machine learning algorithms, spatial models include potential spatial relationships between observations, whose measures are computed as spatial weights matrices. Matrix W is NxN matrix with element w_{ij} as spatial relation measure between observations

I and *j*. Weights can be defined as binary weights where $w_{ij} = 1$ when observations *i* and *j* are neighbours and $w_{ij} = 0$ in other cases, or inverted distance weights where $w_{ij} = \frac{1}{d_{ij}^2}$.

For the purposes of the following paper, all mentioned in this chapter spatial models are computed in two different variants of applied spatial weights matrices: k-Nearest Neighbour Weights Matrix and Inverted Distance Weights Matrix. Among all computed spatial models, Spatial Error Model with k-Nearest Neighbour Matrix is a best fitting model to original data and will be compared in terms of predicting power with other algorithms included in the following research. Spatial Error Model can be formulated as follows:

$$\begin{aligned} y &= \beta X + u \\ u &= \lambda W u + e, \end{aligned} \tag{4}$$

and k-Nearest Neighbour Matrix is formulated as:

$$w_{ij} = \begin{cases} 1, & when\ unit\ i\ is\ one\ of\ the\ k\ nearest\ neighbors\ of\ unit\ j \\ 0, & otherwise. \end{cases}$$

k-Fold Cross-Validation

The common problem when evaluating machine learning algorithms is overfitting that is situation when model performs well on some data (in-sample), but when estimating the model on out-of-sample data errors are much higher. When splitting dataset into in-sample and out-of-sample, number of samples used for training is reduced and results depend on this random partitioning. In order to solve this problem, first we split original dataset into training and test sets and then k-Fold Cross-Validation method was conducted on training data. In this approach dataset is split into k equal-sized sets and k-1 sets are used for training the model and one remaining set is used for validation. Process is repeated k times, so each of k partitions is used as a validation set exactly once. When tuning models, the goal is to achieve such model that variance of performance in k repetitions of k-Fold Cross-Validation is not significant. The final score of the model is an average of the results of each repetition. For purposes of this paper, k-fold cross validation with k equal to 5 is applied. Each of presented algorithms is learned on the same subsets obtained by k-Fold CV from training data and finally evaluated and compared on test data to find the best solution for real estate price prediction problem. The aim of our research is to estimate such models that performance on the test data is not significantly different than average performance on samples obtained by k-Fold Cross-Validation.

Evaluation metrics

In order to evaluate the performance of machine learning methods, some metrics have to be conducted. The idea is to measure the differences between predicted values and observed ones. Among many different ways of measuring the quality of the fit of the models, the most-commonly used measures for regression problems are chosen and presented below.

The mean squared error (MSE) is calculated as an average of squared differences between predicted and observed values and is defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \tag{5}$$

The root mean squared error (RMSE) is a square root of MSE and is formulated as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2} \tag{6}$$

The mean absolute error (MAE) is an average of absolute differences between predicted and observed values and is given by:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|Y_i - \hat{Y}_i\right| \tag{7}$$

The last metric conducted in this paper is the coefficient of determination, known as R^2. One can think, metric is a proportion of the information in the data explained by the model. R^2 is calculated as square of correlation coefficient between predicted and observed values and is defined by:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)}{\sum_{i=1}^{N}(Y_i - \bar{Y})} \tag{8}$$

For each metrics, Y_i is real value for *i*th observation, $\hat{Y}_i$ is predicted value for *i*th observation, $\bar{Y}$ is mean value of Y and N is number of observations. For the purposes of the evaluation, each of above metrics is calculated for k-Fold Cross-Validation that attempted to find best-performed model for each method. Then, metrics calculated on test data were considered as a final score.

4. RESULTS AND MODELS ASSESSMENT

The outcome of the following research are five models that predict prices on real estate secondary market in Warsaw. As it was mentioned above, the evaluation process of each method was conducted using k-Fold Cross-Validation and four different metrics are calculated to compare the performance of the models and choose best fitting for gathered data. Then models are evaluated on test data to examine stability of the estimated methods.

We attempted to specify stable models that perform well on training and validation datasets obtained using 5-Fold Cross-Validation. Table 1. shows estimated values of assessment metrics for each of presented method. Presented values are averages of calculated metrics for five validation samples obtained by Cross-Validation. One can find that eXtreme Gradient Boosting and Random Forest algorithms perform best out of all presented methods in terms of each provided statistics. Mean value of Mean Squared Error for XGB is 1 488 786,46 and mean of MSE for Random Forest is 1 651 891,81. Other methods have much worse values of mean of MSE: 3 425 190,24 for Linear Regression, 3 018 932,49 for Neural Network and 3 449 455,97 for Spatial Error Model.

Table 1 Mean values of evaluation metrics for 5-Fold Cross-Validation.

Method	MSE	RMSE	MAE	R2
Linear Regression	3 425 190,24	1 850,58	1 342,96	56,21%
Neural Networks	3 018 932,49	1 737,04	1 145,14	61,69%
Random Forest	1 651 891,81	1 284,73	792,06	79,19%
Gradient Boosting	**1 488 786,46**	**1 219,58**	**744,47**	**80,97%**
Spatial Error Model	3 449 455,97	1 857,12	1 347,58	55,90%

Source: Own Preparation.

Both tree-based methods are also the best performed in case of other metrics. Gradient Boosting has mean of RMSE equal to 1 219,58 while mean of RMSE for Random Forest is 1 284,73. Mean of MAE for XGB is 744,47 and for RF is 792,06. Finally, proposed eXtreme Gradient Boosting algorithm explains 80,97% of the information in the data and Random Forest explains 79,19% of the variability of the data.

The results above show that both of algorithm based on trees, eXtreme Gradient Boosting and Random Forrest Regression, perform the best out of five proposed methods. Remaining algorithms reach much worse results in terms of proposed evaluation metrics.

Additionally, specific results of the models performance in k-Fold Cross Validation process, are presented in Appendix.

Finally, specified models were evaluated on test sample obtained using *train, validation, test* method. The 5% of observations were randomly draw from original dataset and were not included in process of model's evaluation using k-Fold Cross-Validation. The Table 2. shows evaluation metrics obtained on test data for each proposed algorithm.

Table 2 Evaluation metrics estimated on out-of-sample data.

Method	MSE	RMSE	MAE	R2
Linear Regression	3 540 992,36	1 881,75	1 371,33	57,67%
Neural Networks	3 338 956,27	1 827,28	1 209,25	60,09%
Random Forest	1 777 771,58	1 333,33	788,41	78,75%
Gradient Boosting	1 717 430,05	1 310,51	764,25	79,47%
Spatial Error Model	3 557 779,11	1 886,21	1 367,72	57,47%

Source: Own preparation.

Obtained results show that specified models are stable and perform on test sample as well as on validation samples obtained by Cross-Validation. That allows to consider proposed methods as valuable tools for prediction prices on real estate market.

Spatial effects

The above results present that machine learning algorithms outperform spatial error model in terms of prediction power on out-of-sample data. However, it is necessary to consider spatial effects when modelling spatial data to avoid bias caused by spatial heterogeneity. In order to detect potential existence of that effects, a number of statistics were introduced in relevant studies. Moran's I is widely used metrics that measure global spatial autocorrelation in provided data. One can think that spatial autocorrelation assumes that observations in neighborhood are more similar than distant ones (Kopczewska, 2006). We attempt to estimate Moran's I for dependent variable (PRICE_M2) and residuals of each implemented model obtained on both in-sample and out-of-sample data.

Firstly, we confirmed the existence of spatial autocorrelation of dependent variable in both in-sample and out-of-sample data. This was states on the basis of p-value lower than 0.05 in both cases. Similar results were obtained for linear regression and neural network residuals. This means that both above methods avoid spatial nature of provided data and obtained results

can be assumed as biased. As it was expected Spatial Error Model explains spatial heterogeneity of in-sample data. However, statistics for test sample shows that propagation of certain effects on neighboring units were not identified by SEM in out-of-sample data. This result points out that spatial error model not only performs bad on test data in terms of prediction power, but also has a weak ability to recognize spatial heterogeneity on out-of-sample data. All obtained p-values for Moran's I are presented in ***Table 3***.

Table 3 The p-values obtained by Moran test.

	in-sample	out-of-sample
PRICE_M2	0,0000	0,0000
Linear Regression	0,0000	0,0025
Neural Network	0,0000	0,0002
Random Forest	1,0000	0,3177
eXtreme Gradient Boosting	1,0000	0,1730
Spatial Error Model	0,9992	0,0005

Source: Own preparation.

Surprisingly, above results lead to conclusion that both presented tree-based algorithms can outperform other methods in terms of prediction power as well as explanation of spatial dependencies. Additionally, figures that show spatial distribution of model's residuals on out-of-sample data were presented in Appendix.

5. CONCLUSIONS

In this paper we attempted to investigate a potential of advanced machine learning methods in predicting prices of real estates. We used data on real estate secondary market in Warsaw, Poland in 2018 to evaluate the performance of linear regression, neural network, random forest, xgboost algorithms and spatial error model. The main goal of this paper was to investigate whether it is possible to specify machine learning model that outperform spatial modelling techniques in terms of prediction power. Also potential of above methods to identify spatial dependencies was examined. We applied *train, validation, test* and k-Fold Cross Validation methods to specify and then evaluate all proposed models. The evaluation metrics were used to compare methods and their performance on obtained data. Finally, Moran's I was tested in order to examine model's ability to recognize spatial heterogeneity.

We discovered that tree-based algorithms - random forest and extreme gradient boosting - outperform other introduced methods. This study confirmed that advanced machine learning

methods have higher prediction power than spatial modelling techniques. Additionally, we show that both implemented tree-based methods have ability to avoid bias caused by spatial correlation. The machine learning algorithms allow to solve more complex problems that do not assume linearity of functional form. Furthermore, we managed to specify models that perform well on training and validation data obtained using k-Fold Cross-Validation as well as on out-of-sample dataset. The stability of estimated models allows to consider these methods as valuable tools for prediction prices on real estate market.

In view of the fact that considering spatial effects is important in analysis of real estate market, further research can focus on combination of common machine learning approaches with spatial heterogeneity that is described by spatial weight matrices. The other challenge for research related to spatial modelling is to handle with real-world scale data and scalability of spatial techniques.

References

Basu, S., & Thibodeau, T. G. (1998). *The Journal of Real Estate Finance and Economics, 17*(1), 61-85.

Bency, A. J., Rallapalli, S., Ganti, R. K., Srivatsa, M., & Manjunath, B. S. (2017). Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Bengio, Y., Goodfellow, I., & Courville, A. (2017). *Deep learning*. Massachusetts: MIT Press.

Biau, G. (2012). Analysis of a Random Forests Model. *J. Mach. Learn. Res., 13*, 1063-1095.

Can, A. (1992). Specification and estimation of hedonic housing price models. *Regional Science and Urban Economics,22*(3), 453-474.

Chen, T. & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In B. Krishnapuram, M. Shah, A. J. Smola, C. Aggarwal, D. Shen & R. Rastogi (eds.), *KDD* (p./pp. 785-794), : ACM.

Cohen, J. P., & Coughlin, C. C. (2008). Spatial Hedonic Models Of Airport Noise, Proximity, And Housing Prices*. *Journal of Regional Science, 48*(5), 859-878. doi:10.1111/j.1467-9787.2008.00569.x

Conway, D., Li, C. Q., Wolch, J., Kahle, C., & Jerrett, M. (2008). A Spatial Autocorrelation Approach for Examining the Effects of Urban Greenspace on Residential Property Values. *The Journal of Real Estate Finance and Economics,41*(2), 150-169.

Elhorst, J. P. (2010). Applied Spatial Econometrics: raising the bar, Spatial Economic Analysis, 5(1), 9-28.

Espinoza, E. & Balaguer, J. (2019). Estimating the effects of urban location on social housing: A spatial hedonic approach.

Freund, Y. & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119-139.

Hastie, T., Friedman, J., & Tisbshirani, R. (2017). *The Elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Hong, J., Choi, H., & Kim, W.- sung. (2020). A house price valuation based on the random forest approach: the mass appraisal of residential property in South Korea. *International Journal of Strategic Property Management*, *24*(3), 140-152.

Kopczewska, K. (2006). *Ekonometria i statystyka przestrzenna z wykorzystaniem programu R CRAN*. Warszawa: CeDeWu.

Limsombunchai, V. (2004). House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. *American Journal of Applied Sciences, 1*, 193-201.

McMillen, D. P. (2010). Issues In Spatial Data Analysis. Journal of Regional Science, 50(1), 119-141.

Osland, L. (2010). An Application of Spatial Econometrics in Relation to Hedonic House Price Modeling. *Journal of Real Estate Research*. 32. 289-320.

Peng, Z., Huang, Q., & Han, Y. (2019). Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboost Algorithm. *2019 IEEE 11th International Conference on Advanced Infocomm Technology (ICAIT)*.

Pinkse, J., & Slade, M. E. (2010). The Future Of Spatial Econometrics. *Journal of Regional Science, 50*(1), 103-117.

The Polish Real Estate Guide Edition 2018 Poland - EY. (n.d.). Retrieved October 5, 2019, from https://ey-people.pl/forms/registration.html?docid=32

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, 82(1), 34-55.

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

Song, C., Kwan, M., Song, W., & Zhu, J. (2017). A Comparison between Spatial Econometric Models and Random Forest for Modeling Fire Occurrence. *Sustainability, 9*(5), 819.

Sun, H., Wang, Y., & Li, Q. (2016). The Impact of Subway Lines on Residential Property Values in Tianjin: An Empirical Study Based on Hedonic Pricing Model. *Discrete Dynamics in Nature and Society, 2016*, 1-10.

Wilhelmsson, M. (2002). Spatial Models in Real Estate Economics. *Housing, Theory and Society, 19*(2), 92-101.

Yu, D., Wei, Y. D., & Wu, C. (2007). Modeling Spatial Dimensions of Housing Prices in Milwaukee, WI. *Environment and Planning B: Planning and Design, 34*(6), 1085-1102.

Zhang, R., Du, Q., Geng, J., Liu, B., & Huang, Y. (2015). An improved spatial error model for the mass appraisal of commercial real estate based on spatial analysis: Shenzhen as a case study. *Habitat International, 46*, 196-205.

APPENDIX

Table A.1. MAE values for k-Fold Cross-Validation.

k-Fold	Linear Regression	Neural Networks	Gradient Boosting	Spatial Model	Random Forest
1	1 359,53	1 181,18	738,07	1 366,43	812,28
2	1 349,84	1 183,08	744,90	1 349,63	777,39
3	1 323,63	1 155,29	729,07	1 328,16	786,69
4	1 330,06	1 136,90	728,57	1 336,93	775,21
5	1 351,76	1 069,26	781,74	1 356,72	808,72
Mean	**1 342,96**	**1 145,14**	**744,47**	**1 347,58**	**792,06**

Source: Own preparation.

Table A.2. MSE values for k-Fold Cross-Validation.

k-Fold	Linear Regression	Neural Networks	Gradient Boosting	Spatial Model	Random Forest
1	3 547 634,80	3 086 542,74	1 462 659,14	3 578 568,77	1 781 673,05
2	3 417 636,88	3 176 546,22	1 524 472,90	3 418 762,47	1 602 244,91
3	3 281 073,17	2 873 258,27	1 379 711,47	3 309 352,39	1 614 579,84
4	3 418 041,45	2 830 132,66	1 429 658,08	3 449 348,91	1 521 723,11
5	3 461 564,90	3 128 182,58	1 647 430,73	3 491 247,30	1 739 238,11
Mean	**3 425 190,24**	**3 018 932,49**	**1 488 786,46**	**3 449 455,97**	**1 651 891,81**

Source: Own preparation.

Table A.3. RMSE values for k-Fold Cross-Validation.

k-Fold	Linear Regression	Neural Networks	Gradient Boosting	Spatial Model	Random Forest
1	1 883,52	1 756,86	1 209,40	1 891,71	1 334,79
2	1 848,69	1 782,29	1 234,70	1 848,99	1 265,80
3	1 811,37	1 695,07	1 174,61	1 819,16	1 270,66
4	1 848,79	1 682,30	1 195,68	1 857,24	1 233,58
5	1 860,53	1 768,67	1 283,52	1 868,49	1 318,80
Mean	**1 850,58**	**1 737,04**	**1 219,58**	**1 857,12**	**1 284,73**

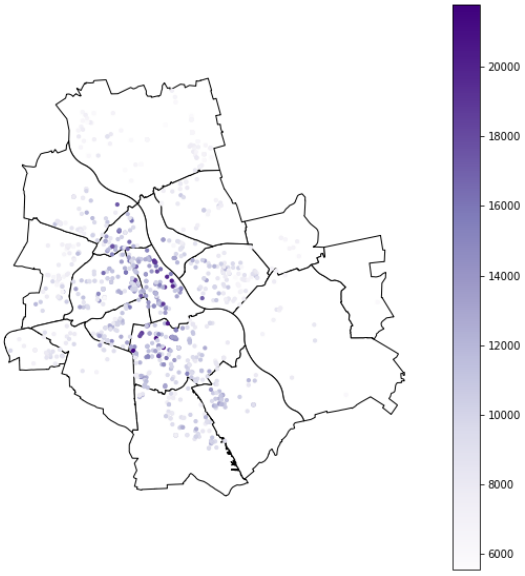
Source: Own preparation.

Table A.4. R2 values for k-Fold Cross-Validation.

k-Fold	Linear Regression	Neural Networks	Gradient Boosting	Spatial Model	Random Forest
1	56,5678%	62,2128%	82,0933%	56,1891%	78,7927%
2	56,0607%	59,1603%	80,4004%	56,0463%	79,5589%
3	56,8842%	62,2432%	81,8696%	56,5126%	79,5274%
4	55,6043%	63,2404%	81,4307%	55,1977%	79,8388%
5	55,9417%	61,6163%	79,0317%	55,5639%	78,2256%
Mean	**56,2118%**	**61,6946%**	**80,9651%**	**55,9019%**	**79,1887%**

Source: Own preparation.

Figure 5 Spatial distribution of price per square meter in out-of-sample data.



Source: Own preparation.

Figure 6 Spatial distribution of linear regression residuals obtained on out-of-sample data.

Source: Own preparation.

Figure 7 Spatial distribution of neural network residuals obtained on out-of-sample data.

Source: Own preparation.

Figure 8 Spatial distribution of random forest residuals obtained on out-of-sample data.

Source: Own preparation.

Figure 9 Spatial distribution of extreme gradient boosting residuals obtained on out-of-sample data.

Source: Own preparation.

Figure 10 Spatial distribution of spatial error model residuals obtained on out-of-sample data.

Source: Own preparation.

University of Warsaw
Faculty of Economic Sciences
44/50 Długa St.
00-241 Warsaw
www.wne.uw.edu.pl