HOUSING PRICES: TESTING MACHINE LEARNING METHODS

Isabelle Chardon

Departamento de Ingeniería Civil, división Transporte

Universidad de Chile

chardoni.ic@gmail.com

Francisco Javier Martínez Concha

Departamento de Ingeniería Civil, división Transporte

Universidad de Chile

fmartine@ing.uchile.cl

Alexandre Bergel

Departamento de Ciencias de la Computación

Universidad de Chile

abergel@dcc.uchile.cl

ABSTRACT

Artificial Neural Networks, Support Vector Machines and Random Forests are applied to the problem of real estate value in Santiago, Chile. The current literature on the use of these algorithms in land use values is limited in quantity and on their assessments, while they seem appropriate to model the complexity of the problem. After comparison of the three algorithms, the conclusion is that the Random Forest is the best for prediction and assess the relative importance of different variables. This algorithm is applied to the entire dataset before differencing it by incomes to conclude on the performance of the model.

Keywords: machine learning, housing prices, importance of variables

1. INTRODUCTION

1.1. Motivation and literature review

Hedonic regression is the most widely tool used for predicting housing prices, because the effects of each attribute on prices can be easily isolated. The coefficients of the regression have direct interpretation as the monetary value of the properties characteristics. Since Rosen (1974), these models also have a theoretical support in urban economics. However, hedonic models are exposed to strong correlation and lack of causal effect, which may imply weak confidence, since the problem of real estate appraisal is complex, nonlinear, and involves many agents. That is, the technique is too simple for the complexity of the system studied. Thus, the lack of flexibility of the hedonic regression justifies considering the use of machine learning algorithms for predicting housing prices as they are capable of dealing with nonlinearities and large datasets. Machine learning is a family of algorithms that have the ability to automatically learn and improve from experience without being explicitly told how to do so. Contrary to analytic approaches, machine learning is able to extract complex patterns for a large number of example data. For example, a machine learning algorithm would be able to distinguish a cat from a dog by discovering patterns from a large number of pictures of cats and dogs. The key aspect here is that at no point one says to the algorithm what a dog or a cat exactly look like. This is exactly what the algorithm has to learn by itself. Machine learning is a prominent application of artificial intelligence that has seen a large adoption thanks to its success in image and text processing.

This paper focuses on investigating three machine learning algorithms: Artificial Neural Network, Support Vector Machines and Random Forests. We will compare these algorithms for a number of reasons: (i) they are proposed by standard scientific libraries such as numpy and scikit-learn and thus easy to use, (ii) they are scalable, robust, and highly tunable algorithms. Such tuning happens using numerous hyperparameters that accompany each of these algorithms.

Artificial Neural Networks (ANNs) for predicting real estate prices were introduced in the early 90s by Borst (1991) as an alternative to the traditional regression based on the hedonic pricing model. An important feature is that, using ANNs allows the modeler to avoid the specification of a functional form relating the attributes of the property to its price, and at the same time taking into account the non-linearity of the causal relations. The results found using ANNs for predicting housing prices in Singapore (Tay and Ho, 1992) and in San Diego (Do and Grudnitski, 1992) suggest that ANNs outperform the traditional regression. However, its use has generated conflicting views, as shown by the results obtained by Worzala *et al.* (1995) after predicting prices of 217 properties. In this study, ANNs are not superior to the multiple regression and inconsistent results between different software packages, between runs of the same package and between run-times are cited. Another recurrent critic of the ANNs is the lack of interpretability due to the high number of hyper parameters involved, which prevents the modeler from isolating each predictor's contribution to price. Nevertheless, this has improved with new specific methods to study the contribution of variables, compiled by Gevrey *et al*. (2003), making it possible to understand the "black box" mechanics of ANNs.

Since Random Forests and Support Vector Machines (SVR) are newer than the ANNs, few examples on real estate pricing can be found in the literature. Nevertheless, studies show promising results. Antipov and Pokryshevskaya (2012) compared eleven methods for predicting residential

prices in Saint Petersburg from a database of 2848 transactions. In particular, they compared Random Forests, ANNs and multiple regression, obtaining mean absolute percentage error (MAPE) of 14,86%, 16,9% and 18,33% respectively. Moreover, results show that the precision of the predictions improves when forecasting the price per square meter instead of the total price of the property. Study by Čeh *et al.* (2018) in Ljubljana (7404 transactions) also found that the Random Forest outperforms the multiple regression (MAPE was 7,27% and 17,28% respectively).

Application of the hedonic model in Santiago, Chile by Figueroa (2012) concludes that attributes such as the density of construction, the size of the property, the socio-economic level in the neighborhood and the presence of a maid's room have a positive impact on the selling price. The corresponding regression model reaches 72,0% of explained variance. Conversely, distance to the historic center (which refers to downtown Santiago), population density and number of rooms have a negative effect. Lavín *et al*. (2011) find that high criminal and contamination rates reduce housing prices. Vega (2017) finds a positive effect on the price of properties located near green spaces, metro stations and bike paths. In contrast, a negative effect is observed for properties located near bus lanes, highways and metro stations when they are constructed in surface. The corresponding explained variance is 76,5%. These studies highlight the diversity of variables perceived by agents as relevant in the location choice, many of them correlated; therefore, it is difficult –or simple impossible– to identify a set of relevant and sufficiently independent variables as required by regression models.

1.2. Description of the algorithms

In this study, SVR, ANNs and Random Forests are applied to perform predictions of housing prices. A brief description of the algorithms is provided below.

Support Vector Machines or SVM (Cortes and Vapnik, 1995), is an algorithm designed to implement linear and non-linear classification by separating clusters of p-dimensional vectors with a (p-1)-dimensional hyperplane. In the nonlinear case, the data is transformed to a higher dimensional feature space to perform a linear separation. The algorithm has been modified for regression by Drucker *et al*. (1997) and is also able to cope with linear and nonlinear problem, in where an implicit mapping (the *kernel trick*) is applied. A detailed description of the SVR model can be found in Smola (2003).

Artificial Neural Networks (ANNs) is a model inspired in the human brain that is able to perform both classification and regression. ANNs can be trained with a training set of observations and then make predictions from a new set of observations. A network is made of layers of artificial neurons, represented by nodes connected to each other by links between layers. Each connection represents a pair of weights that are adjusted during the training phase. The first layer of the network is the *input layer,* a p-dimensional vector of attributes, and the last layer is the *output layer,* where the output is computed. The internal layers, one or more, are *hidden layers.* During the training phase, the network is fed with a set of attributes *(inputs)* and target values or *outputs*; in the case of housing prices prediction the attributes are the characteristics of a property and the target is its price. Each nodal input is the sum of the weighted outputs vectors from the previous layer. An activation function is applied to the nodal input before feeding the following layer with the resulting nodal output. In the *output layer*, the error between the observed target and the network output is computed and then propagated back into the network, where the nodal weights are adjusted in order

to minimize the error. A graphical description of a fully connected neural network with one hidden layer and one output node depicted in Figure 1.

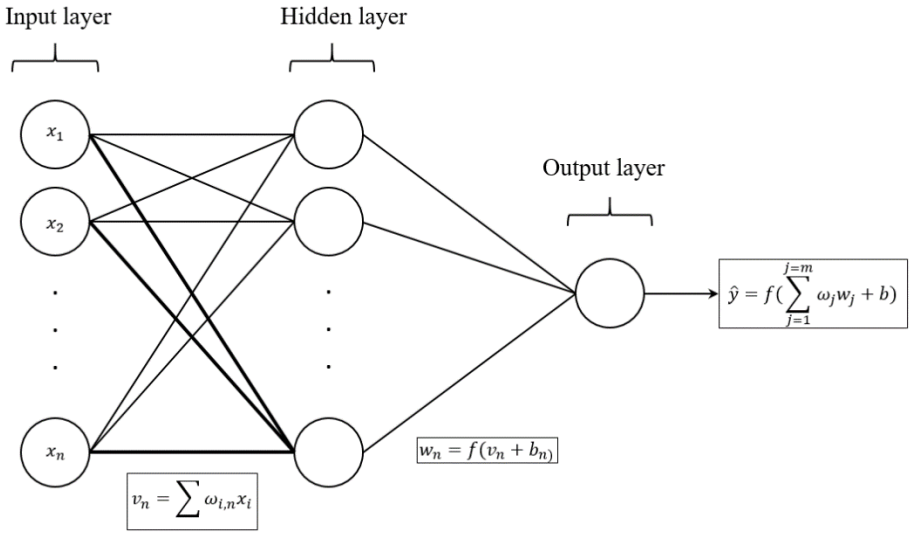

Figure 1: Fully connected ANN with one output layer.
In the hidden layer, each node receives a weighted input v_n, which is subject to an activation function *f* before sending the information signal to the output layer.

The Random Forest model for regression is obtained by combining a multitude of regression trees and outputting the mean prediction of the individual trees. Regression trees are built using the CART algorithm (Classification And Regression Trees; Breimann 2001), which consists in partitioning the features space into distinct and non-overlapping regions. In the CART algorithm, the parameters of the nodes are the variables used for splitting X_i and the splitting value V_j. A metric is computed for every couple (X_i, V_j) in order to select the best parameters. Current metrics are *Gini index*, variance and mean squared error. A graphical representation of a regression tree is depicted in Figure 2.

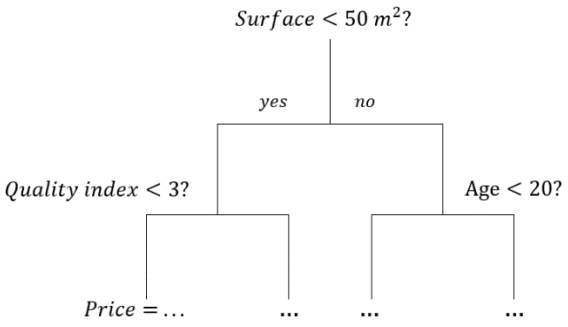

Figure 2: Graphical representation of a regression tree.
In this example, the first splitting variable is *surface* and the splitting value is *50 m^2*.

From the training set, n trees are generated using the bootstrap technique of aggregating. For each tree, a random sample is selected with replacement and is used by the tree to adjust its parameters.

2. THE DATA

The dataset is a sample of 334,353 dwellings transacted over the period 2007-2018 in Santiago, Chile that is taken from a database of 600,602 observations of real estate sales in 33 communes of the Metropolitan Area. The database is property of TocToc.com, a company dedicated to the geo-reference and the valuation of housing properties in Chile. It includes information about internal attributes of the dwelling and external attributes such as access to transportation, health and education services. Each property belongs to a block and to a homogeneous zone in terms of land price; this partition of space is defined by TocToc.com. After having tested the algorithms for predicting the property selling price compared with the selling price per surface unit, the later (in UF/m^2) has been chosen as it generates lower estimation errors. A subset of 205,600 sales was generated to train the algorithms and perform predictions after removing outliers from the database following the rule:

- If the standard deviation of the land price in the corresponding homogeneous zone is in range 0-20 UF/m^2, then the limits are mean of land price +- the standard deviation
- If the standard deviation is in range 20-40 UF/m^2, the limits are mean of land price +- ½ the standard deviation
- If the standard deviation is more than 40 UF/m^2, the limits are mean of land price +- ¼ the standard deviation

The identification of outliers was supervised by a valuation expert from TocToc.com in order to avoid deleting transactions that actually represent market prices. Note that the cleaning was performed before we even initiated our study. It was observed that outliers are usually transactions between members of the same family or transactions of properties whose characteristics do not match with the characteristics registered in the official database.

The following set of variables is available for each property:

- AGE of the property
- SIZE of the dwelling in square meters
- Q_INDEX quality calculated by the Chilean Tax Administration
- NEW dummy variable taking value of 1 if the dwelling was new when sold
- GREEN index that measures the access to green spaces
- DMETRO distance to the nearest metro station
- DSCHOOL distance to the nearest private school
- DCLINIC distance to the nearest private clinic
- DELINQ decil of crime in the neighborhood
- AV_AGE average age of the properties in the neighborhood
- AV_Q_INDEX average quality index of the properties in the neighborhood
- AV_SIMCE average result in the primary education national test in the nearest schools

- INC average income per household in each commune (the administrative zone of Santiago City)
- CBD distance to the city central business district (CBD)
- CEN1 to CEN8 are distances to 8 urban centers
- YEAR of the transaction
- PRICE for the dwelling selling

The statistics of the variables is presented in the appendix, Table A.

The city of Santiago is made of 37 communes with strong concentration of wealthy neighborhoods in the triangle north-east of the CBD. This triangle is characterized by better access to green spaces and private quality health services, better access to public transportation, lower crime rates and larger dwellings with higher quality indexes. Land price is consequently higher than in other communes and prices increase at higher rate, as shown in Figure 3.

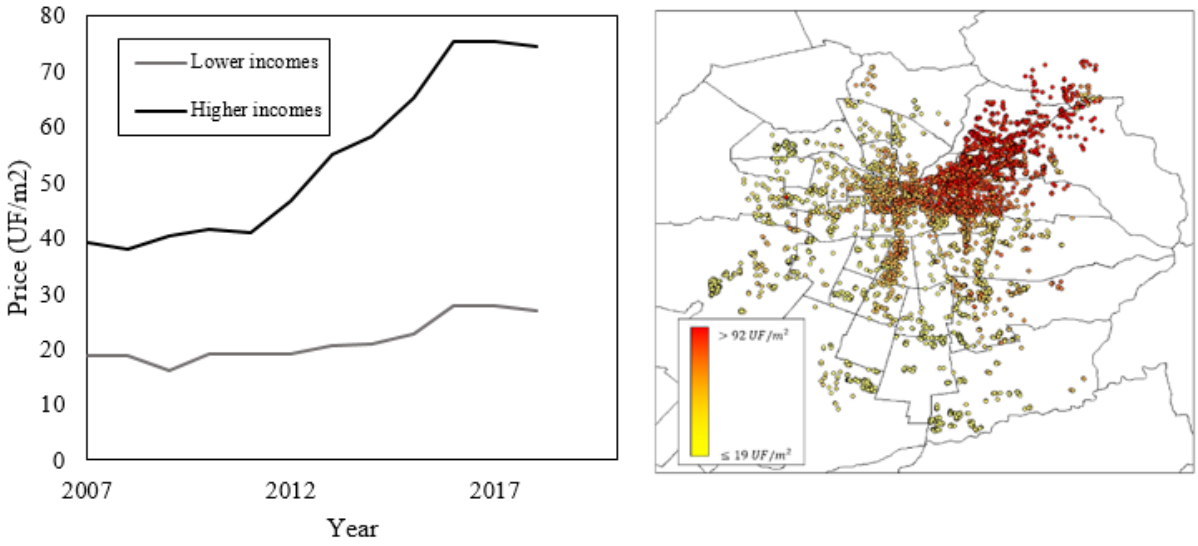

Figure 3: Mean selling price against year of transaction
In lower and higher incomes communes (left) and spatial representation of housing price data in year 2017 (right).

In this paper, the machine learning algorithms have been combined with a grid search algorithm to obtain the best combination of hyperparameters. The resulting models are:

- ANN with five hidden layers and semi linear activation functions. The number of neurons in each layers, from the *input layer* to the *output layer* is 23,25,50,75,1.
- A SVR model with *radial basis function* (rbf) kernel
- A Random Forest model made of 500 unpruned regression trees.

Both Mean Absolute Percentage Error (MAPE) and the percentage of explained variance are adopted as fitting metrics to compare the three methods.
An 80% of the dataset was used to train the algorithm and the remaining 20% was used to perform predictions. Price estimation errors are computed for the training sample and the trained algorithms

are used to forecast the prices in the test sample. At the end of the training procedure, sale prices are estimated by the ANN, the Random Forest and the SVR model. For each algorithm, the results were compared to the actual sales prices computing the MAPE and the coefficient of determination. Pricing errors and percentage of explained variance for data testing and for each algorithm are reported in Table 2. These results show that Random Forest performed the best, followed by the SVR model and the ANN for the MAPE index, while they perform similarly regarding the power to explain the data variance. We conclude that the Random Forest is the most suitable method to explain housing prices in the studied dataset. Moreover, this result is consistent with Antipov and Pokryshevskaya (2012) and Čeh *et al.* (2018) previous studies in very different contexts. Together, these three studies show empirically that Random Forest achieves higher explanatory power than other known machine learning algorithms. Further studies are necessary to confirm the general validity of this preliminary conclusion and to understand the theoretical reason underpinning this result.

Model	Random Forest	ANN	SVR
MAPE (%)	11.27	22.10	14.97
Explained variance (%)	73.7	71.6	67.1

Table 2: Price estimation errors for Random Forest, ANN and SVR.

3. FORECAST AND VARIABLE ASSESSMENT WITH RANDOM FOREST

In this section, the Random Forest model is used for forecasting house prices and assessing variable importance, aimed at obtaining further insights of the use of this algorithm in modeling the housing prices. A second objective is to assess the importance of each variable in the formation of housing prices. Two additional datasets are generated differentiated by the average income per household and commune. The first dataset represents communes with higher incomes (82,375 observations), more than 1.600.000 CLP (Chilean Pesos) per month, and the other dataset represents the other communes (123,225 observations). The average monthly income in the two socioeconomic groups is respectively 2.080.000 CLP and 690.000 CLP, both per month. Predictions and variable importance assessment are also performed for the complete dataset (205,600 observations).

Methods to assess variable importance

Variable importance assessment is performed by computing the increase in node impurity for each variable and by obtaining the distribution of minimal depth in the trees for each variable. Increase in node impurity between a father node N and its child nodes N_l and N_r is based on the *Gini index* and calculated as:

$$\Delta i(N) = i(N) - P_l i(N_l) - P_r i(N_r) \tag{3.1}$$

where i is the *Gini index* and P_l and P_r are the proportions of observations in each child node.
A high increase of the *Gini index* means that the variable used to split the father node induces a reduction in data heterogeneity after splitting.

Partial dependence plots

These plots are generated for most important variable in each case. Partial dependence on variable X_j represents the isolated effect of X_j on the selling price. The corresponding function is defined as:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f(x, x_{iC}) \tag{3.2}$$

where x is the variable for which partial dependence is sought, and x_{iC} is the set of other variables in the dataset.

Results

MAPE and percentage of explained variance obtained for data testing and for the three datasets are shown in Table 3. MAPE are similar for the three datasets and the percentage of explained variance is higher in the communes with high incomes, although the variance of the data in this zone is higher than in low incomes communes for AGE, PRICE, SIZE and GREEN. The larger amount of observations in communes with high incomes can explain the higher value of the explained variance. For each dataset, variable importance is calculated using the increase of node purity and the distribution of minimal depth for each variable. Partial dependence is plotted for the most important attributes.

	Lower incomes	Higher incomes	Total Database
MAPE	10.94	11.31	11.27
Explained variance (%)	72.9	81.1	73.7

Table 3: Pricing error and explained variance for communes with low and high incomes and for the total dataset.

All zones

Distribution of minimal depth and increase of node impurity (Figure 4) and Partial Dependence for all zones (Figure 5) are shown below. Partial Dependence plots depict the selling price per surface unit versus one of the variables X_j while the other variables remain constant.

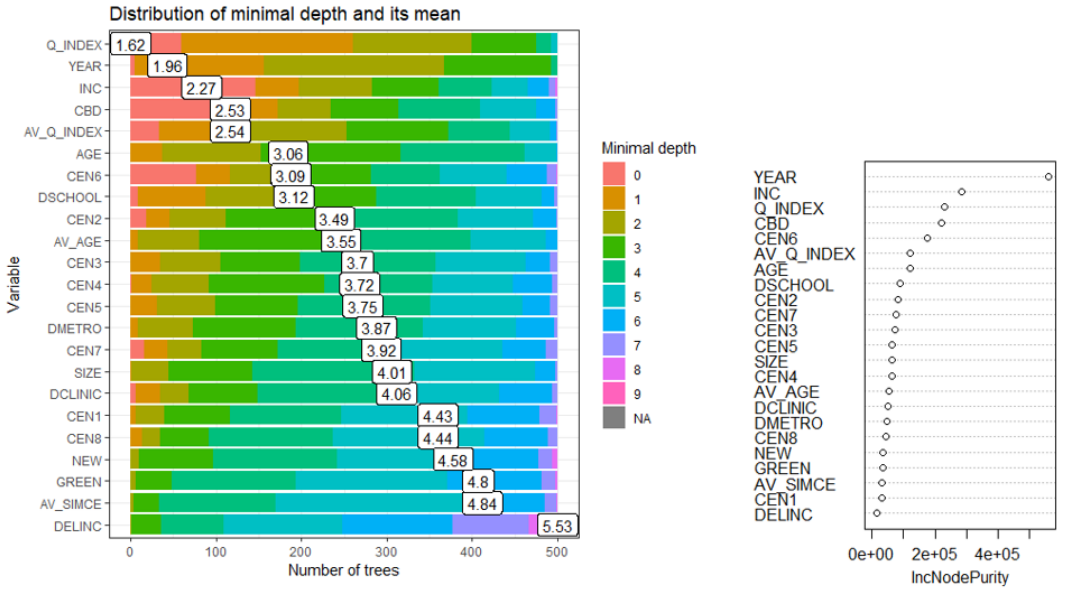

Figure 4: Results for all zones.

Distribution of minimal depth (left) and increase of node impurity (right). The minimal depth of a variable X_j is the level in the trees of the first node split using X_j; the lower the number (see white boxes), the higher the importance of X_j in explaining housing prices.

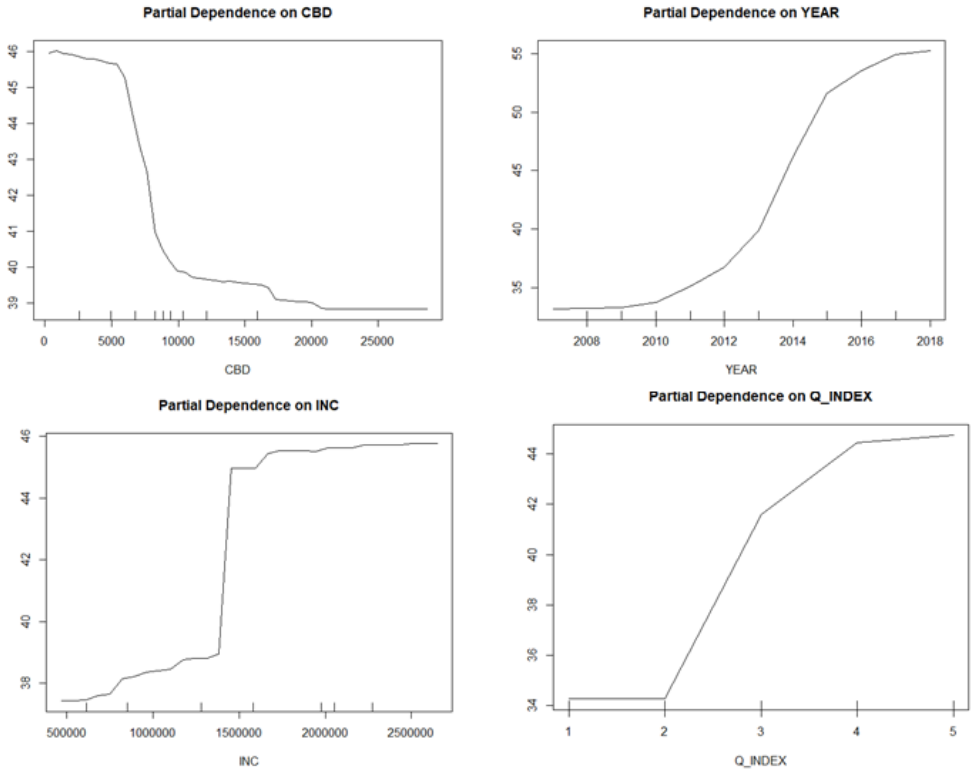

Figure 5: All zones: Partial Dependence

On the distance to CBD (upper left), year of transaction (upper right), mean income per household per commune (lower lefty) and quality index (lower right).

The most important variables for predicting prices in all communes are the year of transaction, the mean income per household per commune, the quality index and the distance to CBD. Partial dependence on the year of transaction shows that the shape of the curve is actually similar to the curves in Figure 3. Partial dependence on distance to CBD and income have similar shapes as these two attributes are highly correlated (communes with high incomes are located near the CBD). According to the first curve, proximity to CBD has a positive impact on selling price within an 8 km radius.

Zones with high incomes

Distribution of minimal depth and increase of node impurity (Figure 6) and Partial Dependence for zones with high incomes (Figure 7) are shown below.

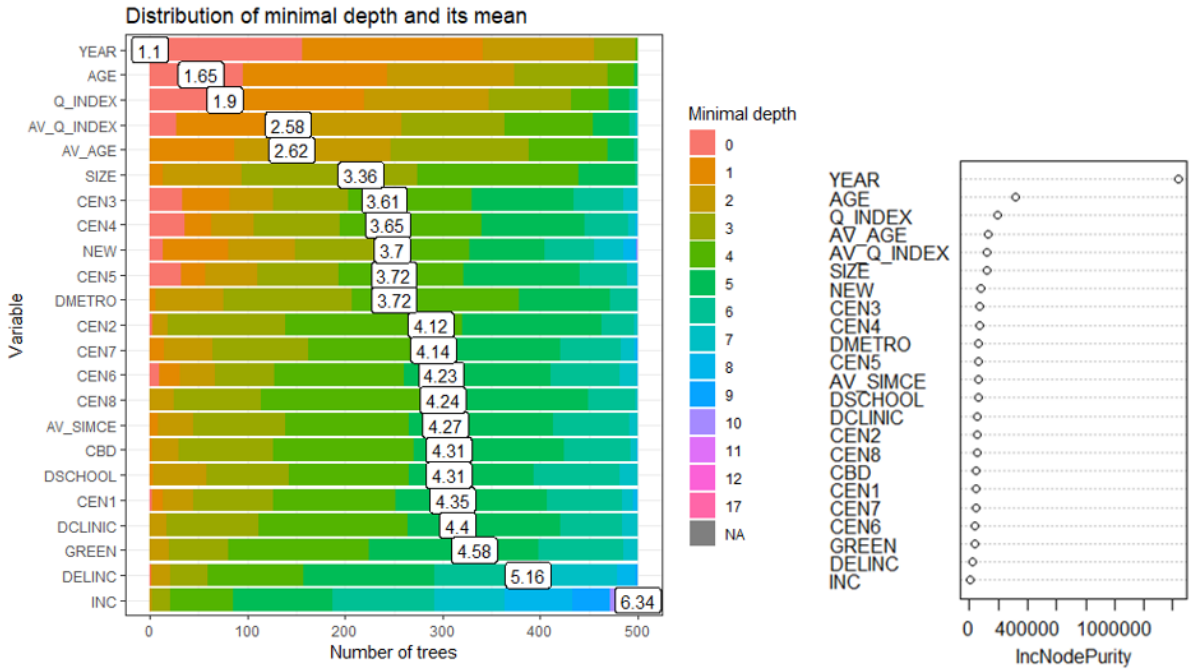

Figure 6: Higher income zones
Distribution of minimal depth (left) and increase of node impurity (right) in the communes with higher incomes.

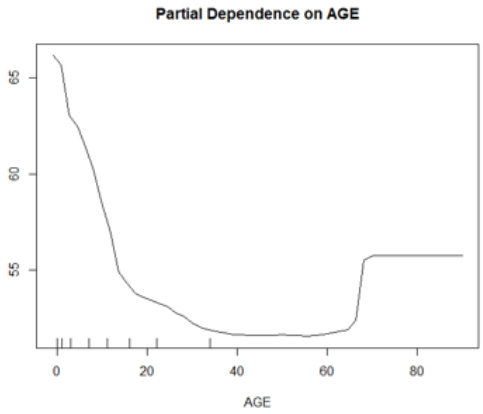

Figure 7: Higher income zones: Partial dependence on age of the dwelling.
Partial dependence on quality index and year of transaction are similar to the curves obtained for the total dataset.

In zones with higher incomes, the most important variable is the year of transaction followed by the age and the quality index of the dwelling. Partial dependence on age of the dwelling shows that prices decrease as the dwellings are older, but this tendency changes after the threshold of 70 years of the dwelling age. An explanation to this result could be the extra value generated by the patrimonial value of the property, which we call the *vintage value.*

Zones with low income

Distribution of minimal depth and increase of node impurity (Figure 8) and Partial Dependence for zones with low incomes (Figure 9) are shown below.

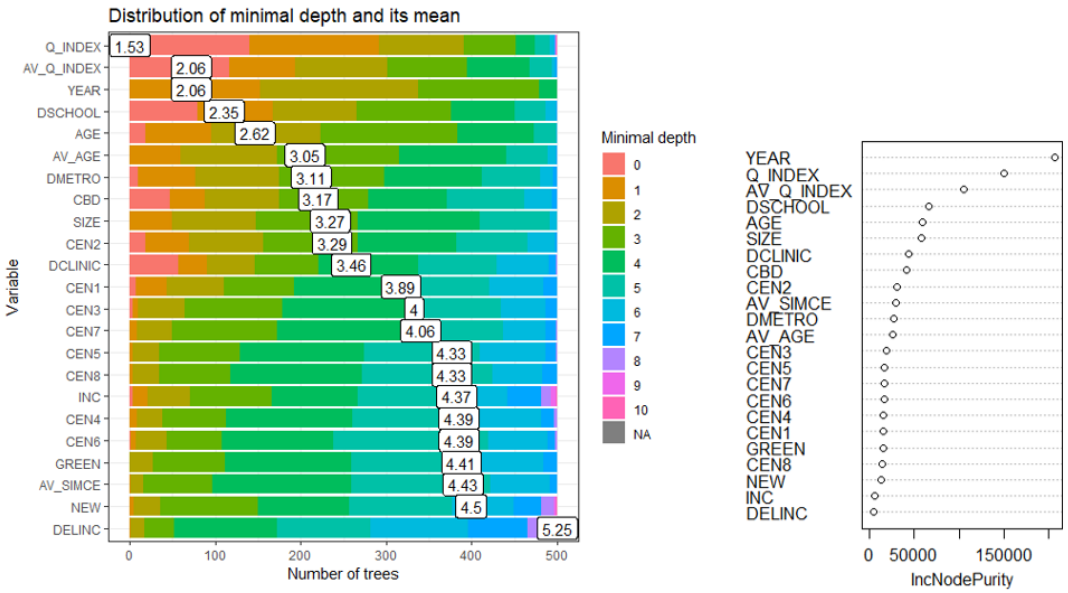

Figure 8: Low income zones
Distribution of minimal depth (left) and increase of node impurity (right) in the communes with lower incomes.

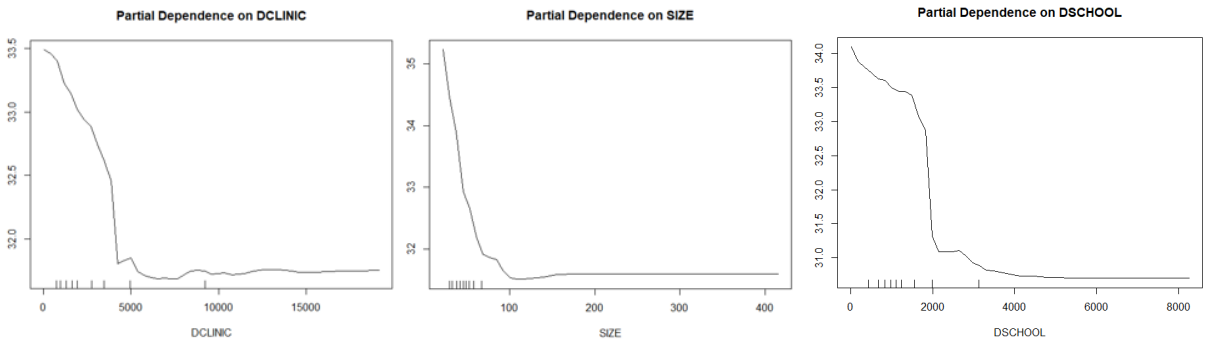

Figure 9: Low income zones: Partial dependence
On the distance to the nearest clinic (left), the size of the property (center) and the distance to the nearest private school (right). Partial dependence on quality index and year of transaction are similar to the curves obtained for the total dataset.

In the zones with lower income, the proximity to private school and private clinics have a positive effect on selling price within a 2 km and 3 km radius respectively. Partial dependence on the size shows that each additional square meter is cheaper as the dwelling size increases.

4. CONCLUSION

This study replicates previous general results regarding the usefulness of machine learning methods in estimating housing prices in urban contexts, because they are more suitable than econometric regressions to handle the complexity of the problem. Such complexity theoretically emerges from the heterogeneity of consumers and their diverse valuation of the physical and socioeconomic environment, which is the underpinning values that theoretical define consumers' willingness to pay at housing auctions, therefore, the housing prices in the market (see Martínez, 2018).

The study provides some useful insights regarding the performance of alternative machine learning algorithms. After applying ANN, Random Forests and SVR to the problem of forecasting housing prices for the case of Santiago longitudinal data, it has been found that the lowest pricing errors are obtained when the dependent variable is the price per square meter instead of the total selling price. Similar result has been found by Antipov and Pokryshevskaya (2012). Additionally, the Random Forests outperformed other algorithms and generated a mean absolute percentage error of 11,27% whereas ANNs and SVR generated an error of 22,10% and 14,97% respectively. The percentage of explained variance is similar between algorithms and to the previous results obtained in Santiago by Vega Flores (2017) and by Figueroa (2012). In addition, the Random Forests has the advantage of limiting overfitting without substantially increasing estimation pricing errors (Breima nn, 2001), while the absence of a functional form between the attributes and the price makes the algorithm flexible and similar to the valuation process based on decision trees made by real estate experts; on the other hand, the absence of a functional form limits an explicit causal-effect interpretation between variables and prices.

We also obtained significant results regarding the importance of specific variables in the construction of housing prices. The *Gini index decrease* and the mean of minimal depth for each variable shows that the year of the transaction, the quality index, the mean income per household per commune and the distance to the CBD are the most important variables. Partial dependence plots also show that the proximity to the CBD has a positive impact on the selling price within an 8 km radius, which matches with the location of the communes with higher incomes. The separation of the dataset in two datasets that represent levels of income allows us to reveal differences between income zones regarding the importance of the variables. In zones with higher income, the most important variables are the year of the transaction and the age of the dwelling whereas in the communes with lower incomes the distance to the nearest private school and the nearest private clinic also have an impact on the selling price. Further studies can be conducted with datasets sorted by dwelling sizes, communes, year of transaction in order to isolate the effects of the attributes on the selling price.

According to the theoretical complexity of the formation of housing prices in urban context described by Martínez (2018) and the empirical results supported by this study, we conclude that the Random Forests algorithm is recommended as a better method than the traditional hedonic regression for predicting housing prices. Analysis of variable importance can be easily made using importance measures and it is possible to obtain partial dependence on each attribute. Finally, the precision of the model can be further improved by adding additional attributes such as the number of rooms, the density of population in the neighborhood and information the purpose of the transaction.

REFERENCE

Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. **Expert Systems with Applications**, *39*(2), 1772-1778.

Borst, R. A. (1991). Artificial neural networks: the next modelling/calibration technology for the assessment community. **Property Tax Journal**, *10*(1), 69-94.

Breiman, L. (2001). Random forests. **Machine learning**, *45*(1), 5-32.

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. **ISPRS International Journal of Geo-Information**, *7*(5), 168.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. **Machine learning**, *20*(3), 273-297.

Do, A. Q., & Grudnitski, G. (1992). A neural network approach to residential property appraisal. **The Real Estate Appraiser**, *58*(3), 38-45.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). Support vector regression machines. ***Advances in neural information processing systems*** (pp. 155-161).

Figueroa, E. (1992). Determinantes del precio de la vivienda en Santiago: Una estimación hedónica. **Estudios de Economía**, *19*(1), 67-84.

Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. **Ecological modelling**, *160*(3), 249-264.

Lavín, F. V., Dresdner, J., & Aguilar, R. (2011). The value of air quality and crime in Chile: a hedonic wage approach. **Environment and Development Economics**, *16*(3), 329-355.

Martínez, F. J. (2018). **Microeconomic modeling in urban science.** Academic Press, Elsevier.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. **Statistics and computing**, *14*(3), 199-222.

Tay, D. P., & Ho, D. K. (1992). Artificial intelligence and the mass appraisal of residential apartments. **Journal of Property Valuation and Investment**, *10*(2), 525-540.

Vega Flores, R. A. (2017). Valoración de amenidades urbanas mediante precios hedónicos: el caso de Santiago de Chile. Magister Thesis, Pontificia Universidad Católica de Chile.

Worzala, E., Lenk, M., & Silva, A. (1995). An exploration of neural networks and its application to real estate valuation. **Journal of Real Estate Research**, *10*(2), 185-201.

APPENDIX

Table A1: Land Use Variables - Statistics

Variable	Min	Max	Mean	Standard Deviation	Unit
AGE	0.0	99.0	8.9	13.7	Years
SIZE	22.0	744.0	61.9	33.3	m^2
Q_INDEX	1.00	5.00	3.24	0.67	-
NEW	1.0	2.0	1.4	0.5	-
GREEN	0.0	320.0	5.1	5.3	m^2/hab
DMETRO	1	29754	1410	1515	m
DSCHOOL	0	8626	1149	1069	m
DCLINIC	18	21428	2770	3122	m
DELINC	1.0	10.0	6.3	2.9	-
AV_AGE	1917.0	2012.0	1993.4	17.7	years
AV_Q_INDEX	1.5	5.0	3.4	0.3	-
AV_SIMCE	200.0	322.7	270.1	16.5	-
INC	468284	2649750	1466092	613560	CLP
CBD	40	50848	9237	5252	m
CEN1	34	43785	6402	4528	m

CEN2	88	48371	7121	4600	m
CEN3	187	37078	10985	3921	m
CEN4	131	35633	20530	4540	m
CEN5	275	34544	19458	4493	m
CEN6	147	38742	19884	5023	m
CEN7	583	38984	16468	4846	m
CEN8	3713	56013	16817	4523	m
YEAR	2007.0	2018.0	2012.4	3.4	Year
PRICE	2.4	2229.4	41.5	21.5	UF/m^2(1 UF =27,931 CLP, july 5, 2019)