Financial vs traditional pricing models: which model explains better real estate prices in Bogotá?

Autor

Juan Camilo Salgado Ramírez

Trabajo presentado como requisito para optar por el título de Magíster en Economía

Director

Jesús Gilberto Otero Cardona

Facultad de Economía

Maestría en Economía

Universidad del Rosario

Bogotá - Colombia

2022

FINANCIAL VS TRADITIONAL PRICING MODELS: WHICH MODEL EXPLAINS BETTER REAL ESTATE PRICES IN BOGOTÁ?

JUAN CAMILO SALGADO RAMÍREZ

May 2022

Abstract.

This paper estimates two property valuation models and determines which one is more accurate in predicting real estate prices in Bogotá. One method considers that the price of a house is explained by its relationship with properties with similar characteristics, while the other method is implemented from the financial literature (Gordon (1959)) and proposes that a property is a financial asset that generates periodical income through rent. By using a novel database of online publications of properties in 2020 in Bogotá, I found that the financial valuation method has a similar predictive capability on real estate prices compared to more traditional methodologies. These results allow the possibility to implement more robust financial valuation techniques to the real estate market.

I thank my advisor Jesús Otero for his valuable comments and supervision and *Habi* for providing me access to its database on real estate properties in Bogotá.

1. Introduction

The real estate sector is an important segment of the Colombian economy. According to official data from Departamento Administrativo Nacional de Estadística (2022), this sector represented 9.07% of Colombian GDP in 2019. If the value of construction of residential properties is also considered, this share rises to 12%. Also, financial institutions in Colombia held a total of 1.2 million residential loans, which represents 3% of total Colombian population of 44 million people (Departamento Administrativo Nacional de Estadística (2019)).

In Bogotá –Colombia's capital– there are 7 million inhabitants –16% of Colombia's population– in around 2.3 million residential properties, according to Departamento Administrativo Nacional de Estadística (2018). It has also been estimated that around 120.000 properties are transacted every year (La Galería Inmobiliaria (2019)). Given this scenario, understanding how property prices are settled is an interesting question for research.

In almost all transactions, prices are settled using information about similar properties that have been sold in a particular zone of the city. Therefore, properties are considered as any other good, in which the price is determined by its characteristics. But what if properties are considered as financial assets, in which their price would be determined as the present value of some expected returns?

With that objective in mind this work arises: this paper aims to determine which type of valuation method better describes the price generation mechanism that exists in real world property data. For this task I implement two different methodologies to price a property. First, I follow a *traditional methodology*, which is the most used method to price a property and has a common pattern: all properties that have similar characteristics –area, bedrooms, location, etc– must have similar prices. Then, I implement a standard method following the existing financial literature on asset valuation and adapt it to the particularities of the real estate market. Therefore, analyzing the

performance of the *financial methodology* applied to the real estate market relative to the more *traditional methods* is the objective of this paper and the novelty of my work.

The *traditional methodology* is comprised by different methods that vary in complexity and precision. For example, some methods are manual in the sense that a person reviews by hand all houses close in distance to another one and extract the properties that have similar characteristics. Then, the price of the property in interest can be computed as the mean or median of the prices from that set of similar properties. This method is known in Colombia as *peritajes* and is vastly used by banks to determine the value of a house before giving a mortgage, or by local governments to estimate the value of all housing units in a given city and then determine the value of tax per house.

The existing literature on this methodology implements more robust statistical methods and also machine learning algorithims. One key model used in this literature is the hedonic price model. Its first use can be traced back to Waugh (1928), who analyzed the relationship between the attributes of different kinds of vegetables on their price. He found that, for example, a greener color in asparagus is associated with a higher price. In real estate, several studies have used this approach to estimate the relationship between changes in property prices and changes in amenities. For example, Sirmans et al. (2006) performed a meta regression analysis using 58 studies that implemented an hedonic price methodology to determine the stability of the coefficients of the main variables and found that the estimated coefficients for some characteristics vary significantly by geographical location; Lynch and Rasmussen (2001) found that houses located in the top two crime deciles neighbourhoods are selled for 39% less than a comparable house in other areas; Evans (2012) exploits the exogenous variation of a 1.7 billion USD loss in a local government treasure to analyze its effects on the real estate market and found a 1.64 - 3.2 billion loss in market value; or Clark and Cosgrove (1990) developed a two-stage hedonic model to estimate the value of public safety and found that a 10% increase in their metric of public safety is associated with

an increase in monthly rents by 1.3%. In Colombia, following a hedonic price methodology Cabrera-Rodríguez, Mariño-Montaña, and Quicazán-Moreno (2019) estimate a price index for new housing units in Bogotá using data from *La Galería Inmobiliaria*. The authors complemented the hedonic model with spatial analysis and obtained a higher level of accuracy on the price index relative to other studies that only used stratification or hedonic models without the use of the location of the property. Given that in the literature of the real estate market the use of hedonic price models is widely used, I decided to carry out an estimation following this approach.

Another group of works approaches the problem using machine learning algorithms. The primary objective of the studies that use this methodology is prediction rather than inference. As Mullainathan and Spiess (2017) outline:

> "Machine learning [...] evolves around the problem of prediction: produce predictions of y from x. The appeal of machine learning is that it manages to uncover generalizable patterns. In fact, the success of machine learning at intelligence tasks is largely due to its ability to discover complex structure that was not specified in advance. It manages to fit complex and very flexible functional forms to the data without simply overfitting; it finds functions that work well out-of-sample. Many economic applications, instead, revolve around parameter estimation: produce good estimates of parameters β that underlie the relationship between y and x. It is important to recognize that machine learning algorithms are not built for this purpose." (P. 88)

Therefore, the use of machine learning models should be preferred to applications where finding a precise $\hat{y}$ is the main objective rather than understanding how a certain attribute or an exogenous shock affects an outcome variable. One clear example of a problem of this type is the work by Krizhevsky, Sutskever, and Hinton (2012). In this study, the idea was to correctly classify 1.2 million high-resolution images into 1000

different categories. For this task, the authors used a large, deep convolutional neural network with 60 million parameters and 650.000 neurons. The model was trained with images previously categorized by human labelers and performance was tested out of sample. The authors achieved an error rate of 15.3%, which placed them top 5 in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC)-2012 competition. In this scenario, understanding how a particular pixel affects the probability of being in a particular category of image is not relevant. The most important thing in this case is finding a model that accurately fits the data and correctly labels each image in its real category.

In the real estate literature, multiple studies have used machine learning techniques and obtained a higher level of accuracy on price predictions than more standard methods. Examples include the work by Mullainathan and Spiess (2017), where they fitted an OLS regression and a random forest to predict the log-dollar house value of owner-occupied units in the 2011 American Housing survey. They obtained an R^2 of 41.7% for the OLS regression and an R^2 of 45.5% for the random forest, both measured out of sample; the study by Trawiński et al. (2017), where the authors compare the performance of machine learning models, linear regression and *expert algorithms*, which are the standard methods used by real estate appraisers licensed professionals. They found that machine learning algorithms had the lower error of prediction, followed by the OLS regression and then by the *expert algorithms*. In both studies, the potential of the usage of machine learning techniques to predict real estate prices with a high degree of accuracy was stated.

The machine learning literature also compares the performance and accuracy of different types of models. For example, Trawiński et al. (2017) fitted 11 different machine learning models to predict academic outcomes from a dataset of student attributes. The authors found that random forest and decision tree algorithms were the most accurate models with an out-of-sample accuracy of 96.71% and 97.03% respectively. For the

real estate market, Jha et al. (2020) estimated 4 machine learning algorithms –namely logistic regression, random forest, voting classifier, and XGBoost– to predict real estate prices and found that the XGBoost algorithm presented the highest accuracy rate at 92%; Truong et al. (2020) predicted house prices in Beijing using random forests, extreme gradient boosting, light gradient boosting and ensemble models –which are constructed as the combination of multiple machine learning models– and despite they found that the ensemble models had the lower percentage error, all algorithms had a similar performance –all errors ranged from 16.35% to 16.94%–[1]. Since it is not clear in the literature that a certain machine learning model is the best to predict house prices, I decided to estimate an XGBoost algorithm because its performance is of the highest in all the papers reviewed.

Despite the apparent existing conflict between the two approaches, it should also be noted that combining both types of methods –namely OLS regression models and machine learning techniques– have the potential to produce good results both in terms of inference and prediction. For example, in Colombia Pérez-Rave, Correa-Morales, and González-Echavarría (2019) use both hedonic pricing models and decision trees with advanced sampling methods to predict house prices and found improved performance when combining both types of methods. In my current work, however, I will only estimate an hedonic price model and a machine learning approach using an XGBoosting algorithm[2] separately to test performance on house prices prediction[3].

From the financial literature, the first theoretical framework for asset pricing can be traced back to the Capital Asset Pricing Model (CAPM) developed by Sharpe (1964). In this paper, the author develops a theory to determine an optimal capital market equilibrium given investors' preferences and a risk-free rate of return. From this model,

[1]For a detailed information about the performance of the multiple models implemented in the literature, see Table 5.
[2]For a more detailed discussion of the algorithm see subsection 3.1.2
[3]I decided not to combine them to understand how each method performs independently on real estate prices prediction.

in equilibrium, the return of a particular asset i will be higher than the return of a risk-free asset if and only if that asset i is positively correlated with the market and the market is expected to have a higher return than the risk-free asset.

This theory, however, is not best suited to analyze the dynamics of the real estate sector. For example, Claus (2013) states that the CAPM is focused on one-period analysis whereas the real estate investing strategy is a multi-period problem. Therefore using the CAPM as the theoretical foundation is not a good choice. A better financial model to analyze the dynamics of the real estate market is the model proposed by Gordon (1959). This model considers the case of a stock that generates a certain dividend that grows at a constant rate in perpetuity. In this scenario, the price of that stock will be given by the present value of the infinite series of future dividends [4]. Despite that this model was mainly proposed to explain prices of stocks, it can also be adapted to the real estate sector because both markets share common characteristics. For example, a property generates periodical income through monthly rent and that rent grows over time. Therefore, I will use this financial valuation method to estimate the price of property prices in Bogotá.

For my estimations I will use data for both selling and renting properties published in the main online listings platforms of Colombia using web scraping techniques. This approach was previously used by Cárdenas Rubio, Chaux Guzmán, and Otero (2019) for different Colombian cities and showed the potential of this information to estimate the evolution of the price of properties over time and across geographical locations. In my current work, however, I will only use information for Bogotá.

Given the theoretical framework previously explained about the traditional and financial valuation methods, the main contribution of my work is the empirical estimation of financial pricing techniques in the real estate sector; predictions are then

[4]For a detailed discussion, see subsection 3.2

contrasted with estimations of traditional valuation methods and I show that the financial valuation method offers similar performance to the traditional method. This result allows the possibility to use financial valuation methods when information is scarce and more traditional methodologies cannot be used.

Finally, this thesis is organized as follows: a *data* section which explains all the sources of information that I use in this paper, how I obtained them, the main data processing techniques I used and their implications, a *methodology* section that explains the methods I use for the *traditional valuation* and *financial valuation* as well as how the performance of each method will be measured, a *results* section that discusses the main findings from both models, a *conclusion* section that summarizes the main findings of this paper, a *complementary data* section which contains additional information and a *appendix* section which contains the main tables and results.

2. Data

To measure both traditional and financial valuation models I have access to a rich dataset composed by the online listings of properties published on two of the most used real estate websites in Colombia –Fincaraiz and Ciencuadras–. According to Figure 3, in December 2020 *Fincaraiz* and *Ciencuadras* accounted for a total of 300k searches on Google, which placed them top 3 in the most used listing services in Colombia. From this dataset I observe the main characteristics of a property: sale price, rent price, area, stratum[5], parking garages, years built, elevators, bathrooms, bedrooms, floor level, address, business type, latitude and longitude –Summary statistics for each variable are presented in Tables 3 and 4–. This dataset is composed by properties published either for sale or for rent –to avoid potential problems I eliminated properties published as both business types–. This dataset was created by periodically running a code script to extract the information from the websites. An example of how the

[5]This is an official classification system created in Colombia to determine the socio-economic level of a house given it's characteristics, surroundings, etc.

information is published in each web portal can be seen in Figures 4 to 9. Given that the period of extraction was only for 2020 –pre and during Covid–, results could be sensible to the potential effects that the Covid pandemic had on the real estate sector. Particularly, Covid could have affected the intrinsic relationship between the attributes of a house and it's price. For example, given the expansion of work-from-home policies implemented by companies, demand for houses close to workplaces could contract, reducing prices for those properties, while demand for properties with better ammenities and more space in farthest zones of the city could increase, rising the price of properties in those places. Therefore, my estimations are local to the Covid-era and should not be generalizable to any period of time.

To complement this dataset I used information about the amenities close to each property according to Google maps services. The Google maps API works by searching for a particular word and it returns the locations of the sites that contain that word. Therefore I searched for words to identify the number of bus stops, shopping malls, hospitals and schools in Bogotá city. Then I computed the total amount of each site in a radius of 500 meters around each property[6].

2.1. **Cleaning process.**

Because properties are published directly by each house owner and mistakes can be made when typing information, the dataset is not fully clean. To partially solve this problem I restricted the sample to observations that fall within a reasonable range of the main variables: sale price between 30 MM COP - 10.000 MM COP or rent price of 200.000 COP - 15 MM COP, constructed area between 10 mt2 to 2000 mt2, 1 to 5 bathrooms, 1 to 6 bedrooms, 0 to 4 parking garages, 0 to 4 number of elevators. A property with any value outside these ranges is considered an outlier and is not considered in the analysis. Each of these ranges represented at least 90% of total observations. This

[6]The Google maps API allows a certain amount of requests for free per day. For that reason I limited the searches for those sites of interest

database also had missing values on multiple variables of interest. To avoid potential problems when inputting missing values, I decided to drop observations with at least one missing value in a variable of interest.

Additionally, given the methodology I use in financial model that is explained in Subsection 3.2, I implemented a process to clean the addresses of the publications. That process uses regular expressions to identify all patterns from the main parts of an address. For example, if an address contains any of the words *transv*, *tranv*, *tv* or *tr* I assume that those addresses come from a *Transversal* and label them as *T*. Therefore for all the possible combinations of addresses in Colombia, which are *Calle*, *Carrera*, *Transversal* or *Diagonal* I constructed a dictionary of all the possible values and label them as *C*, *K*, *R* or *D* respectively. Finally, I cleaned the rest of the address by extracting the numbers and their companion letters.

The relevance of this process is clear in the following example. Suppose there are two a addresses written as *Avenida Carrera 26 55c - 37sur* and *Av. K 26 55c - 37sur*. Given the difference in the way the addresses were registered, a simple match will produce a mistake. Therefore the algorithm will return the same for both properties *K+26+55c+37sur*, which solves the problem. By using this standardization process I can merge properties located in the same building that are published for rent and for sale.

After this cleaning process I obtained a total of 26K for rental properties and 19K for sale properties. Table 3 presents the summary statistics of properties published for rent and Table 4 of the properties published for sale. One interesting fact from these tables is that the characteristics of the properties published as rent are similar to the characteristics of the properties published as sell. This is relevant because it allows that both sources of information can be merged because are similar.

Finally, Figure 10 shows the spatial distribution of ask prices in the city and Table 2 the median of the most relevant variables by locality. Figure 10 shows that the most

expensive properties are located mainly in the northeastern quadrant and according to Table 2 the most expensive localities are *Chapinero* and *Usaquen* with a median sale price of 830 million COP and 495 million COP respectively.

3. Methodology

3.1. **Traditional valuation.**

3.1.1. *Hedonic price estimation.*

The objective of the hedonic price model is to determine how many each attribute of a property contributes to its total price. Therefore, the price of a house can be decomposed as the sum of the individual values of each of its characteristics. Equation 3.1 represents the estimation I will perform:

$$P_i = \beta_0 + X_i'\Upsilon + \epsilon_i, \tag{3.1}$$

where P_i is the price in million pesos per square meter (COP/m^2) of the property i. X_i is a set of the characteristic of the house[7] and the error term is ϵ_i.

3.1.2. *Gradient boosting algorithm.*

The gradient boosting algorithm is an *additive model*, which consists of the creation of a robust model using simple models and adding them in the process (Cook (2016)). In this case the simple models are decision trees, which use the different covariates to split the data using simple rules to assign a predicted value to each group[8]. The key parameters for this algorithm can be divided into two groups: the parameters of the tree models and the parameters of the gradient boosting. For the tree model I just

[7]The variables in X_i include: area, area^2, stratum, garages, years built, elevators, bathrooms, bedrooms, floor level and the number of bus stops (this includes normal buses stops as well as *Transmilenio* stations), shopping malls, hospitals, schools in a radius of 500 mts.

[8]For example, in a tree where there exists only one covariate, gender, a simple rule would be to split the data between men and women and take the average of the variable of interest for each group.

modified the max depth of the tree, which controls how many rules one can create in each tree. For the gradient boosting parameters I modified the number of trees to be generated and the learning rate, which is a weight that restricts the contribution of each tree to the whole model. I also used as a loss function the squared of the error term. For choosing the best of these parameters I performed a grid search, which is a process that creates a matrix of the desired parameters, estimates multiple models using all combinations of parameters and selects the parameters that have the best performance[9]. For this estimation I used the XGBoost algorithm, which is more accurate than the gradient boosting algorithm because it uses regularization techniques to control over-fitting of the model. The same set of property characteristics used in the hedonic price model, X_i, will be used here to generate the trees.

3.2. **Financial valuation.**

This method consists mainly of estimating the price of a property as the discounted cash flows of future renting earnings. For this purpose I consider a property as a financial asset known as perpetuity, which is an asset that generates a periodical source of income that lasts forever. In theory, a property can be rented forever. However one could argue that once the property reaches a certain age it cannot be rented more because it has reached its useful life. To avoid overcomplications of this exercise I assume that properties can indeed be rented forever as a perpetuity.

According to Gordon (1959), the traditional way to valuate a particular perpetuity is this:

$$PV = \frac{cashflow}{discountrate}, \tag{3.2}$$

[9]In this case, a max_depth = 3, n_estimators = 300 and a learning_rate = 0.3 were the best parameters in the grid search. Therefore I used them in the estimations.

where PV is the present value of the perpetuity, $cashflow$ is the expected payment of the perpetuity and the $discountrate$ is the rate at which that cashflow will be discounted.

This formula cannot be applied directly to the real estate sector because a property has some particular characteristics that make the simple use of the formula in equation 3.2 more difficult. For example, a property is expected to have its rent increased over time. Another problem is that the property will not be rented at all times, or that there exist some associated costs of having a property –for example, periodical maintenance or payment of taxes–. Given these complexities, the formula from equation 3.2 can be modified to take those problems into account as follows:

$$PV_i = \frac{(monthly\_rent_i * 12 * occupancy\_rate_i) - associated\_costs_i}{discount\_rate_i - growth\_rate_i}, \tag{3.3}$$

where PV_i is the present value of property i, $monthly\_rent_i$ is the total amount of rent that a particular property i obtains in a particular month, 12 is the number of months in a year, $occupancy\_rate_i$ is the percentage of time that a property i would be rented in a particular year, $associated\_costs_i$ represents all the associated costs of having a property, $discount\_rate_i$ is the rate at which all the future payments will be discounted and $growth\_rate_i$ is the growth rate of the yearly payments from rent. The most important task in this particular model is to estimate all those parameters accurately.

For the variable $monthly\_rent_i$ I first standardized all the addresses using the method outlined in Section 2. Then, I computed the median of the rent price for each standarized address and merged those values with each property that is for sale[10]. That means that I have information about the price at which a particular property is sold

[10]To be more precise, I measured the median of the rent value per square meter, to allow differences between rent values as prices rise.

as well as the median rent value of properties in that same building. $monthly\_rent_i$ will therefore be the median price of rent of all properties located in the same address of property i. By multiplying it by 12 I estimate how much rent a particular property i is expected to generate in a year.

The variable $occupancy\_rate_i$ is more difficult to estimate mainly due to the abscence of data. In an ideal scenario, this could be estimated easily by having a database of all properties listed for rent and being able to identify which of these properties were indeed rented. However, I can only observe all properties that are offered for rent at a given moment, and cannot identify which of those properties were rented. For that reason, I decided to use the estimation by the global real estate consultant Colliers InternationalGroup (2019) that put this occupancy rate at around 90%[11].

The variable $associated\_costs_i$ is driven mainly by two forces: maintenance and taxes. From these two categories, taxes are the principal contributor to the cost of having a house –of course, in some cases a major issue could occur in a house and a high value of maintenance is required, but this is not true on average–. For this reason, I consider only the cost of taxes. In Figure 11 I show that, by using a sample of 20 properties bought by *Habi*[12] and comparing the median rent of houses in each building and the value of taxes paid for each property, approximately the annual value paid as property taxes equates to one month of rent. For that reason, $associated\_costs_i$ will be equal to $monthly\_rent_i$[13].

For the $discount\_rate_i$ parameter I used the average capital cost for the acquisition of new properties during 2019 and 2020 (Figure 12), which is 10%. The idea behind

[11]In their report, they do not measure the occupancy rate for residential real estate but rather for the offices market. Therefore one assumption that I make is that both sectors and comparable and the occupancy are similar

[12]*Habi* is a company that buys and sells residential properties in Colombia and Mexico.

[13]Due to company restrictions I just present a sample of 20 properties. However, in some analysis performed using more data, the relationship of one month of rent per property to pay taxes holds.

this choice is that an investor would only invest in buying a property if the expected return he makes is higher than the capital cost of buying the house.

And finally, the $growth\_rate_i$ is regulated by the Law 820 of 2003 in Colombia, which specifies that the annual increment of the monthly rent value should not exceed the inflation rate of the previous year. Therefore $growth\_rate_i$ is determined by the forecast of inflation. In this context, using the target of inflation proposed by Colombia's Central Bank is the best choice. According to the Bank's official policy (Banco de La República (2022)):

> "The inflation target has been set at 3% by the Bank's Board of Directors –with a permissible deviation of ±1 percentage point–. This target refers to consumer price inflation, which is measured statistically as the annual variation in the Consumer Price Index (CPI)." (P. 1)

For that reason, I will use 3% for the $growth\_rate_i$ parameter.

Replacing all those parameters in equation 3.3 results in the following formula:

$$PV_i = \frac{monthly\_rent_i * (12 * 90\% - 1)}{10\% - 3\%}, \tag{3.4}$$

The formula in equation 3.4 is the one I will use to estimate the financial model.

3.3. **Comparison of performance.**

Given that the financial valuation method restricts the estimation to a subsample that has information for both the initial price at the launch of the project and future closing data, I will compare predicted prices for both methods relative to the ask prices of this subsample.

The main performance metric I will use is the *Mean Average Percentage Error* (MAPE) which is defined as follows:

$$MAPE_m = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y_{tm}}}{Y_t}\right|, \tag{3.5}$$

where $MAPE_m$ is the MAPE of the model m, n is the number of observations, Y_t is the real price of the property t and $\hat{Y_{tm}}$ is the predicted price for the property t using model m. The lower the MAPE the better the performance of the model.

Finally, Each estimation will be trained using $\frac{2}{3}$ of the dataset and tested out of sample in the other $\frac{1}{3}$ of the dataset.

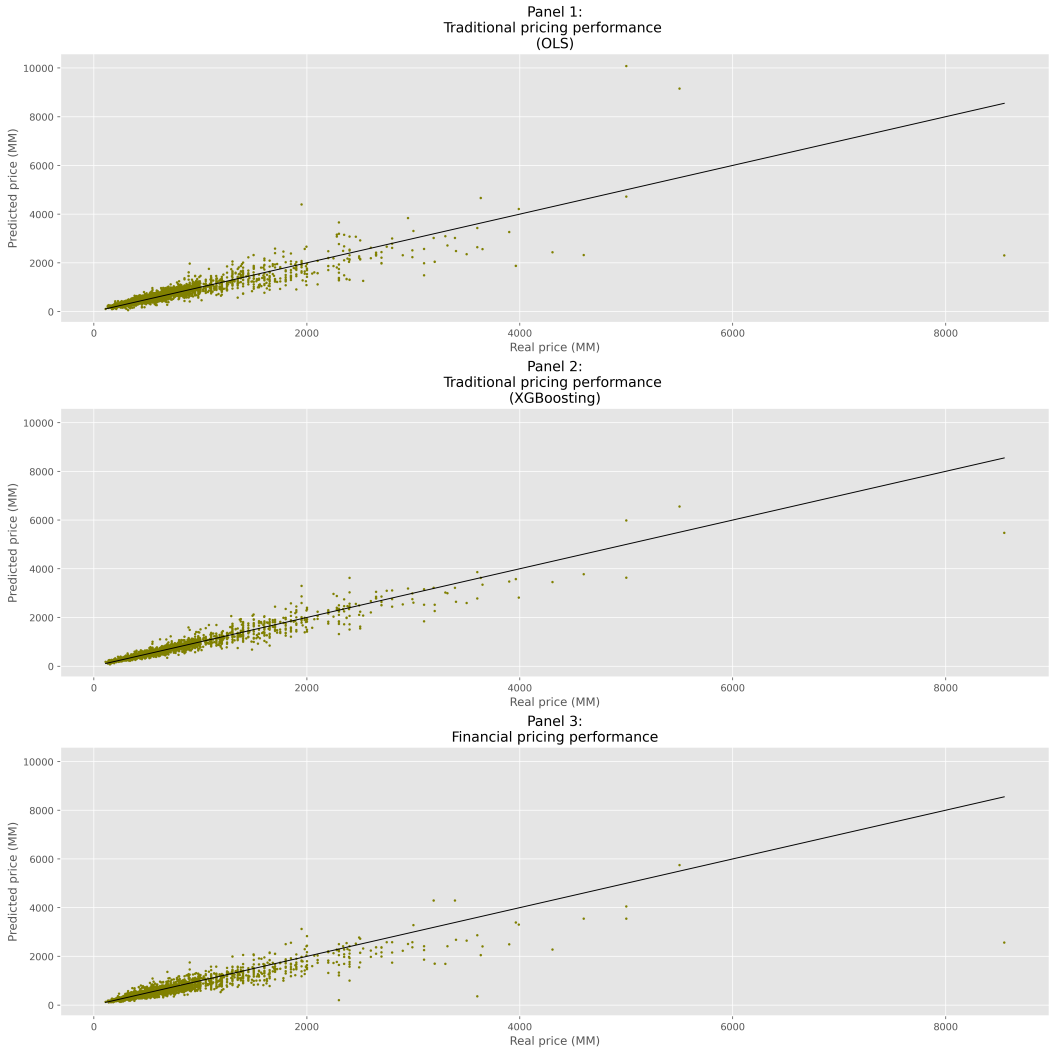
4. Results

Figure 1 shows the adjustment of both the traditional models and the financial model:

For each graph, the y-axis represents the predicted price of the model and the x-axis is the real price of the property –prices are in millions–. Therefore, each point represents the tuple of the predicted price and the real price for each property. In a perfectly fitted model, those two values should be the same. For that reason, the farthest each point is from the 45-degree line, the worst the performance of the model. However, given that there exists a lower density of expensive properties than cheaper ones –most properties have a sale value of less than 1000 million COP–, extracting conclusions about fitting and performance from this figure is not that clear.

One way to solve this problem is to plot observations that have a sale price of less than some particular value. By observing the descriptive statistics of the for-sale properties in Table 4, 75% of properties have a sale value of less than 698 MM COP.

FIGURE 1. **Adjustment of all models**



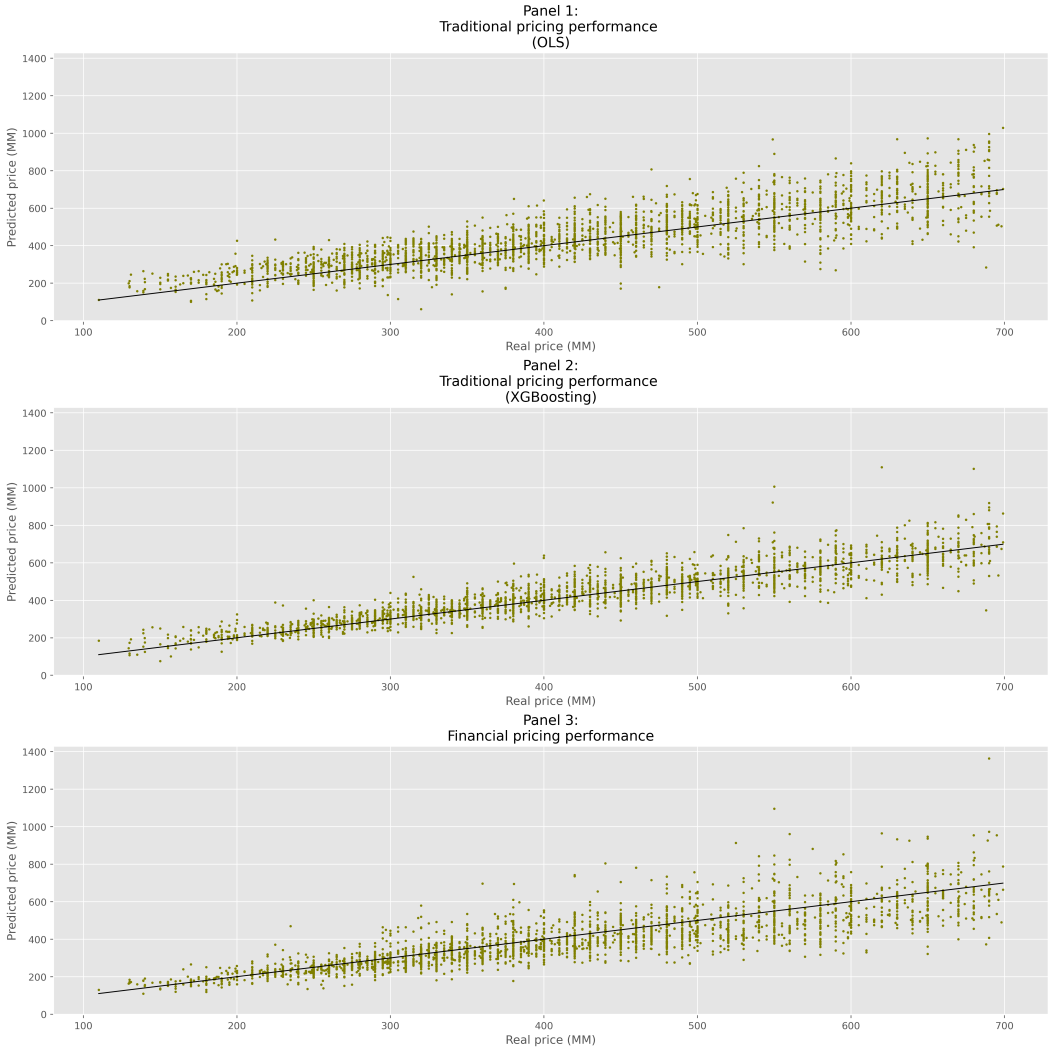
R^2 of 0.59, 0.69 and 0.57 respectively

Therefore I will plot the results for properties with a price of less than 700 MM COP in Figure 2 [14].

Figure 2 facilitates the interpretation of the results. First, it should be noted that all models seems to have a decent predictive capability: the predicted price follows an ascending trend through the identity line. This means that, on average, all models

[14]I only make this to allow better visualization of the results. Therefore I do not restrict the sample and reestimate each model, I just plot the results of properties with a sale value of less than 700 MM COP.

FIGURE 2. **Adjustment of all models with a sale price $<$ 700 MM**



R^2 of 0.63, 0.76 and 0.61 respectively

capture the intrinsic reasons that make an expensive property expensive or a cheaper one cheap.

However, for a particular sale price results are not that precise. For example, consider all properties with a sale value of 400 MM COP in Panel 1 of Figure 2. Predicted values for these properties range from 250 MM COP to 620 MM COP. This range is huge and shows the variance of the predictions for any given property.

It also seems that variance increases with sale price: for cheaper properties, the distance between the points and the 45-degree line is smaller than the same distance of expensive properties. This fact could be explained by two reasons: either the models cannot explain expensive properties, or this just shows that, for a particular percentage error, the absolute difference between predicted price and real price is higher for an expensive property than a cheaper property. In other words, consider two properties, one with a sale value of 100 MM COP and another property with a sale value of 500 MM COP. If all the models tend to perform with an average error of 10%, the most likely prediction price for the 100 MM COP property will lie within the interval 90 - 110 MM COP, but for the expensive property the range would be 450 - 550 MM COP. Therefore the distribution of points of the cheaper properties would seem to be closer to the identity line than the distribution of the expensive properties, but this should not mean that the models predict better cheaper properties. It just shows that for a high valued property, a fixed percentage error generates a larger interval than for a cheaper one.

Now, to analyze the performance of each of the models and to complement the visual interpretation of the results, I consider relevant to analyze the MAPE of each of the models. The following table shows these results for both Figure 1 and Figure 2:

TABLE 1. **Mean Average Percentage Errors for the models**

	All properties	**Properties with price < 700MM**
Traditional pricing (OLS)	15.84%	15.47%
Traditional pricing (XGBoosting)	10.37%	9.75%
Financial pricing	14.74%	13.76%

In terms of predictive performance and as it was discussed in Section 3.3, the best model will be the one with the smaller MAPE. Given this condition, it is clear that the traditional pricing model that uses XGBoosting outperforms the other two models because its MAPE is lower for both samples. This fact shows that Machine Learning

algorithms have the capability to improve the accuracy of any model due to their ability to capture more complex relationships and patterns in the data.

Another relevant conclusion from this table is that the MAPE of each model in the restricted sample is smaller. This means that the models perform better for those properties that have a smaller price. This result could be explained by the fact that the density of points with higher prices is low, and therefore fitting a model with those observations is problematic.

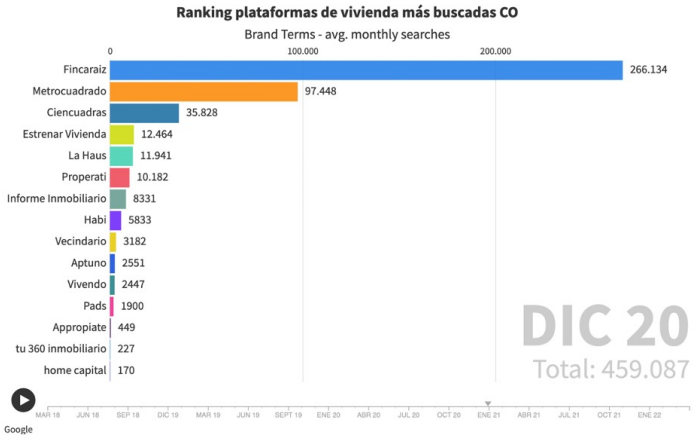
But the most important conclusion that can be extracted from Table 1 is that the financial pricing model has a pretty good performance capability. It even outperforms the traditional pricing model estimated using OLS in both samples. This fact shows that thinking in a property as an asset and following a standard financial valuation is indeed an acceptable way to price a house. However, given that its MAPE is relatively high, using this type of model to predict exactly to the cent the price of a property is not the most accurate way. In that case, given these results, following a more traditional approach that implements the use of machine learning algorithms should be the preferred choice.

5. Conclusion

In this paper I examined which price generation mechanism is more precise to explain the formation of real estate prices. For this task I used a rich database coming from online properties listed on two of the main websites in Colombia in 2020 and estimated two models –an hedonic regression and a XGboosting algorithm– that can be classified as *traditional models* because consider that the price of a property is determined by its relationship with properties that have similar characteristics. Then, I compared the results with a model used in the financial literature which evaluates a house as a financial asset that generates perpetual returns in the form of rent and grows over time. I found that the performance of the XGboosting is the best of all models, followed by the *financial model* and then by the hedonic regression. These results suggest that valuating a property as a financial asset is indeed a good choice and allows the possibility to implement different financial techniques in the real estate market, especially if information about comparables homes is scarce in a particular location –for example, in new constructed buildings is easier to find information about the rent price of a property rather than the sale price because historical information just does not exist–. Improvements to this work include more robust estimations of the parameters of the financial model, the inclusion of more relevant variables in the traditional methodology –for example, crime related data–, or even the implementation of machine learning techniques in the financial modelling of real estate prices to increase accuracy and performance.

6. Complementary Data

Figure 3. **Total searches of real estate listing services in Dec, 2020**



Source: Peña Talero (2022)

Figure 4. **Example of a listed property in Fincaraiz**

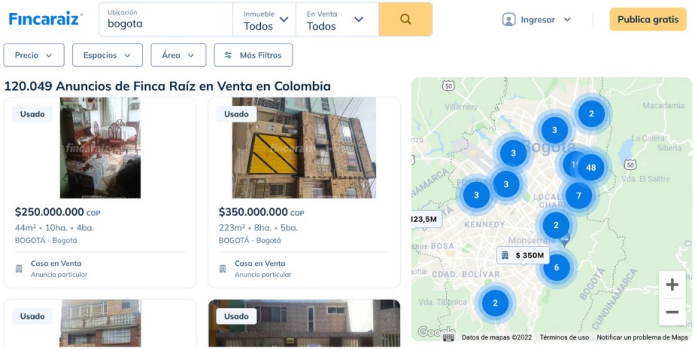


FIGURE 5. **Example of a listed property in Fincaraiz**

FIGURE 6. **Example of a listed property in Fincaraiz**

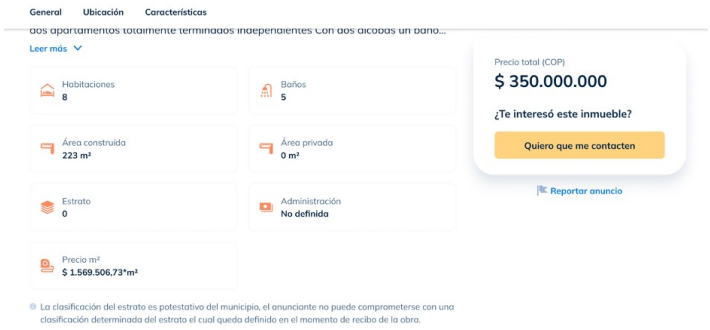


FIGURE 7. **Example of a listed property in Ciencuadras**

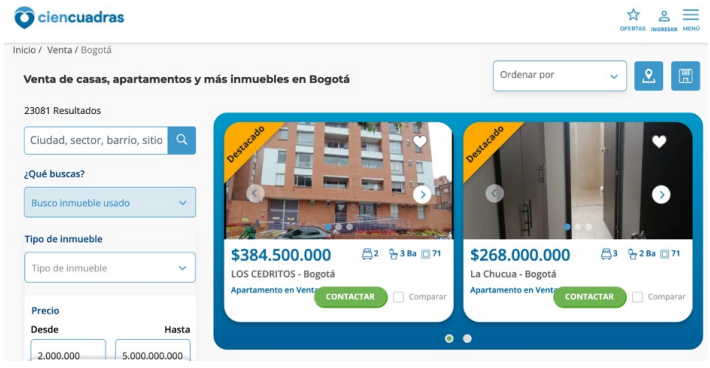


FIGURE 8. **Example of a listed property in Ciencuadras**

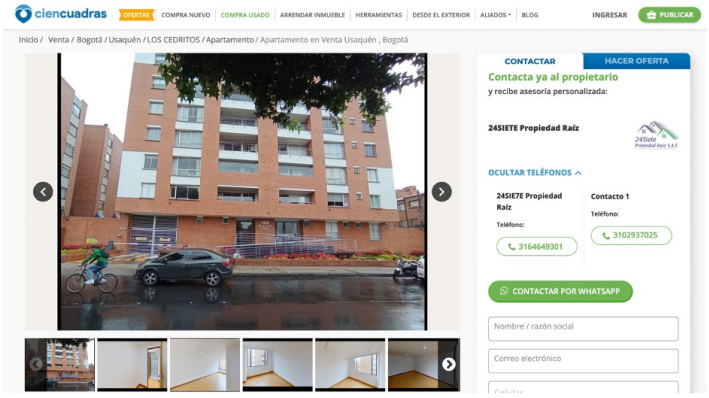


FIGURE 9. **Example of a listed property in Ciencuadras**

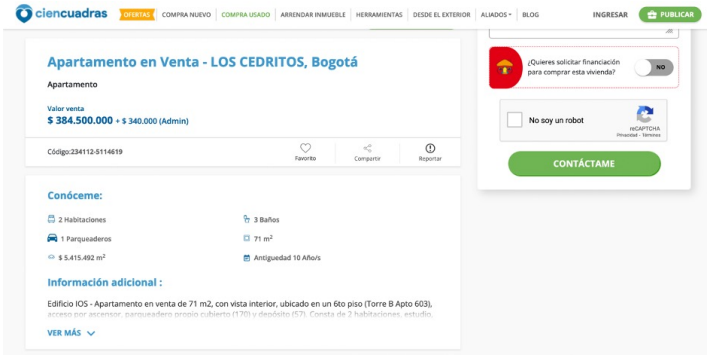
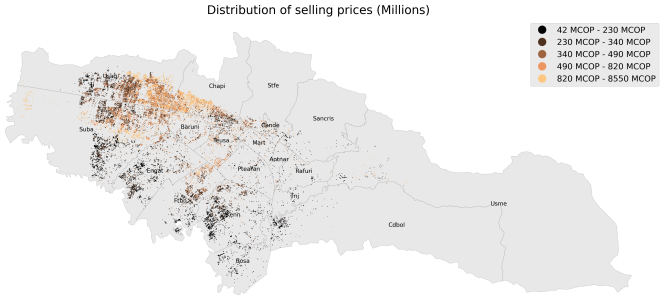


FIGURE 10. **Distribution of ask prices**



Antonio Nariño (Antnar), Puente Aranda (Ptearan), Usaquen (Usaq), Tunjuelito (Tnj), Teusaquillo (Teusa), Suba (Suba), Santa Fe (Stfe), San Cristobal (Sancris), Rafael Uribe Uribe (Rafuri), Los Martires (Mart), Barrios Unidos (Baruni), Kennedy (Kenn), Fontibon (Ftbn), Engativa (Engat), Ciudad Bolivar (Cdbol), Chapinero (Chapi), Candelaria (Cande), Bosa (Bosa), Usme (Usme)

FIGURE 11. **Rent price vs tax paid**

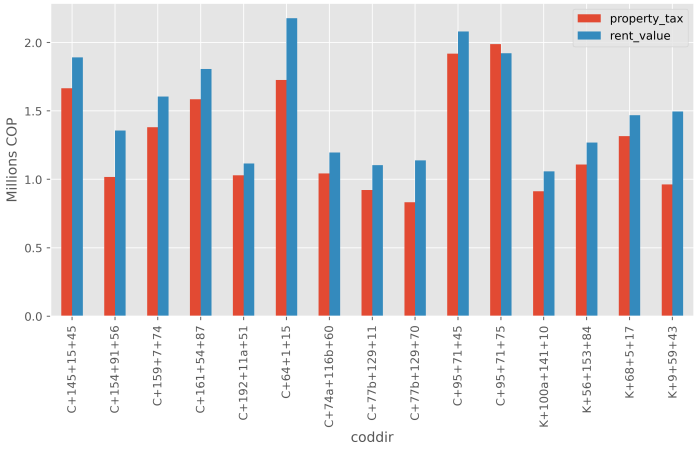
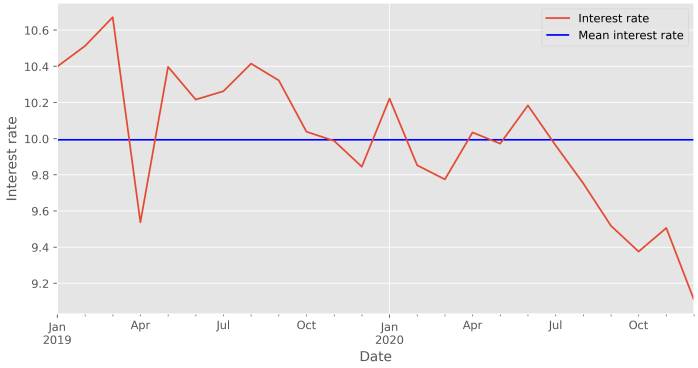


FIGURE 12. **Historical interest rates for mortgages offered by commercial banks**



Source: Central Bank of Colombia

7. Appendix

Table 2. **Median of each variable by locality of Bogotá**

Locality	**Price (MM)**	**Area**	**Bedrooms**	**Bathrooms**	**Years built**
Antonio Nariño	295.0	84.00	3.0	2.0	25.0
Barrios Unidos	450.0	85.00	3.0	2.0	19.0
Bosa	120.0	48.00	3.0	1.0	13.0
Candelaria	375.0	79.00	3.0	2.0	24.0
Chapinero	830.0	120.00	2.0	3.0	23.0
Ciudad Bolivar	164.4	55.00	3.0	2.0	10.0
Engativa	205.0	59.16	3.0	2.0	23.0
Fontibon	319.0	70.00	3.0	2.0	17.0
Kennedy	185.0	57.00	3.0	2.0	14.0
Los Martires	440.0	86.00	3.0	2.0	16.0
Puente Aranda	270.0	61.00	3.0	2.0	17.0
Rafael Uribe Uribe	180.0	55.00	3.0	2.0	11.0
San Cristobal	168.0	52.00	3.0	2.0	8.0
Santa Fe	435.0	68.00	2.0	2.0	13.0
Suba	390.0	85.00	3.0	2.0	17.0
Teusaquillo	420.0	77.89	3.0	2.0	20.0
Tunjuelito	190.0	57.81	3.0	2.0	27.5
Usaquen	495.0	95.00	3.0	3.0	17.0
Usme	260.0	68.00	3.0	2.0	12.0

TABLE 3. **Summary statistics of the properties for rent**

	Rent (MM)	**Area**	**Elevators**	**Bedrooms**	**Bathrooms**
count	26341.00	26341.00	26341.0	26341.00	26341.00
mean	2.76	93.53	1.0	2.25	2.28
std	2.11	58.97	0.0	0.89	0.91
min	0.37	13.00	1.0	1.00	1.00
25%	1.50	58.00	1.0	1.00	2.00
50%	2.00	75.00	1.0	2.00	2.00
75%	3.07	109.00	1.0	3.00	3.00
max	14.90	1667.00	1.0	6.00	5.00

	Floor level	**Garages**	**Stratum**	**Years built**
count	18207.00	26341.00	26340.00	11378.00
mean	5.01	1.47	4.77	17.30
std	3.10	0.66	1.05	10.67
min	1.00	1.00	1.00	2.00
25%	3.00	1.00	4.00	9.00
50%	4.00	1.00	5.00	14.00
75%	6.00	2.00	6.00	26.00
max	16.00	4.00	6.00	62.00

TABLE 4. **Summary statistics of the properties for sale**

	Price (MM)	Price m2 (MM)	Area	Elevators	Bedrooms
count	19034.00	19034.00	19034.00	19034.0	19034.0
mean	671.20	5.86	109.68	1.0	2.6
std	551.33	1.80	63.00	0.0	0.8
min	81.00	1.40	20.00	1.0	1.0
25%	350.00	4.62	68.00	1.0	2.0
50%	490.00	5.52	90.00	1.0	3.0
75%	780.00	6.70	130.00	1.0	3.0
max	8550.00	24.43	624.00	1.0	6.0

	Bathrooms	Floor level	Garages	Stratum	Years built
count	19034.00	14394.00	19034.00	19033.00	14812.00
mean	2.59	5.39	1.67	4.76	16.43
std	0.94	3.44	0.73	1.02	10.24
min	1.00	1.00	1.00	1.00	2.00
25%	2.00	3.00	1.00	4.00	9.00
50%	2.00	5.00	2.00	5.00	13.00
75%	3.00	7.00	2.00	6.00	24.00
max	5.00	16.00	4.00	6.00	62.00

	Bus stops	Shopping malls	Hospitals	Schools
count	19016.00	19016.00	19016.00	19016.0
mean	2.43	2.47	2.14	4.0
std	7.22	2.23	2.63	3.6
min	0.00	0.00	0.00	0.0
25%	0.00	0.00	0.00	2.0
50%	0.00	2.00	1.00	4.0
75%	3.00	4.00	3.00	5.0
max	79.00	15.00	10.00	23.0

TABLE 5. **Goodness of fit of different models used in the literature**

	Metric	OLS	Neural Network	Decision Tree
Trawiński et al. (2017)	MAPE	82.18%	82.79%	97.03%
Jha et al. (2020)	MAPE	–	–	–
Truong et al. (2020)	RMSLE	–	–	–

	Metric	Random Forest	Logistic regression	XGBoost
Trawiński et al. (2017)	MAPE	96.71%	–	–
Jha et al. (2020)	MAPE	89.5%	85%	92%
Truong et al. (2020)	RMSLE	16.57%	–	16.6%

	Metric	Ensemble models
Trawiński et al. (2017)	MAPE	–
Jha et al. (2020)	MAPE	–
Truong et al. (2020)	RMSLE	16.35%

MAPE: Mean Absolute Percentage Error; RMSLE: Root Mean Squared Logaritmic Error

References

Banco de La República, . (2022). "Colombia's inflation targeting strategy". In: URL: https://www.banrep.gov.co/en/monetary-policy.

Cabrera-Rodríguez, Wilmar Alexander, Juan Sebastián Mariño-Montaña, and Carlos Andrés Quicazán-Moreno (2019). "Modelos hedónicos con efectos espaciales: una aproximación al cálculo de índices de precios de vivienda para Bogotá". In: *Borradores de Economía* 1072.

Cárdenas Rubio, Jeisson Arley, Francisco José Chaux Guzmán, and Jesús Otero (2019). "Una base de datos de precios y características de vivienda en Colombia con información de Internet". In: *Revista De Economía Del Rosario* 22(1), pp. 75–100. URL: https://doi.org/10.12804/revistas.urosario.edu.co/economia/a.7768.

Clark, David E. and James C. Cosgrove (1990). "HEDONIC PRICES, IDENTIFICATION, AND THE DEMAND FOR PUBLIC SAFETY*". In: *Journal of Regional Science* 30(1), pp. 105–121. DOI: https://doi.org/10.1111/j.1467-9787.1990.tb00083.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9787.1990.tb00083.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9787.1990.tb00083.x.

Claus, Munk (2013). *Financial Asset Pricing Theory*. Vol. 1st ed. OUP Oxford. ISBN: 9780199585496.

Cook, Darren (2016). *Practical Machine Learning with H2O*. O'Reilly Media, Inc.

Departamento Administrativo Nacional de Estadística, DANE (2018). "CNPV 2018 VIHOPE v2 [Data file]". In: URL: https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/censo-nacional-de-poblacion-y-vivenda-2018.

Departamento Administrativo Nacional de Estadística, DANE (2019). "anexos cartera IVtrim19 [Data file]". In: URL: https://www.dane.gov.co/index.php/estadisticas-

por-tema/construccion/cartera-hipotecaria-de-vivienda/cartera-hipotecaria-de-vivienda-chv-historicos.

Departamento Administrativo Nacional de Estadística, DANE (2022). "Anexos produccion constantes IV 2021 [Data file]". In: URL: https://www.dane.gov.co/index.php/estadisticas-por-tema/cuentas-nacionales/cuentas-nacionales-trimestrales/pib-informacion-tecnica.

Evans, Thomas A. (2012). "An Estimate of the Accuracy of Hedonic Real Estate Valuations Using the Orange County Bankruptcy". In: *Economica* 79(316), pp. 703–720. ISSN: 00130427, 14680335. URL: http://www.jstor.org/stable/23274788.

Gordon, M. J. (1959). "Dividends, Earnings, and Stock Prices". In: *The Review of Economics and Statistics* 41(2), pp. 99–105. ISSN: 00346535, 15309142. URL: http://www.jstor.org/stable/1927792.

Group, Oxford Business (2019). *Colombian real estate market shows beginning of a revival of sustainable growth*. URL: https://oxfordbusinessgroup.com/overview/policymakers-and-private-developers-are-steadily-laying-groundwork-sustainable-growth-raising.

Jha, Shashi Bhushan et al. (2020). "Machine Learning Approaches to Real Estate Market Prediction Problem: A Case Study". In: *CoRR* abs/2008.09922. arXiv: 2008.09922. URL: https://arxiv.org/abs/2008.09922.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

La Galería Inmobiliaria, . (2019). "Informe de Mercado 2019". In: *Publicaciones propias de La Galería Inmobiliaria*.

Lynch, Allen K. and David W. Rasmussen (2001). "Measuring the impact of crime on house prices". In: *Applied Economics* 33(15), pp. 1981–1989. DOI: 10.1080/00036840110021735. eprint: https://doi.org/10.1080/00036840110021735. URL: https://doi.org/10.1080/00036840110021735.

Mullainathan, Sendhil and Jann Spiess (2017). "Machine Learning: An Applied Econometric Approach". In: *The Journal of Economic Perspectives* 31(2), pp. 87–106. ISSN: 08953309. URL: http://www.jstor.org/stable/44235000.

Peña Talero, Sergey (2022). "Ranking de plataformas de vivienda en Colombia". In: URL: https://public.flourish.studio/visualisation/9334685/?utm_source=linkedin.

Pérez-Rave, Jorge Iván, Juan Carlos Correa-Morales, and Fabián González-Echavarría (2019). "A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes". In: *Journal of Property Research* 36(1).

Sharpe, William F. (1964). "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk". In: *The Journal of Finance* 19(3), pp. 425–442. ISSN: 00221082, 15406261. URL: http://www.jstor.org/stable/2977928.

Sirmans, G. Stacy et al. (2006). "The Value of Housing Characteristics: A Meta Analysis." In: *Journal of Real Estate Finance and Economics* 33(3), pp. 215–240. ISSN: 08955638.

Trawiński, Bogdan et al. (2017). "Comparison of expert algorithms with machine learning models for real estate appraisal". In: *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pp. 51–54. DOI: 10.1109/INISTA.2017.8001131.

Truong, Quang et al. (2020). "Housing Price Prediction via Improved Machine Learning Techniques". In: *Procedia Computer Science* 174. 2019 International Conference on Identification, Information and Knowledge in the Internet of Things, pp. 433–442.

ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2020.06.111. URL: https://www.sciencedirect.com/science/article/pii/S1877050920316318.

Waugh, Frederick V. (1928). "Quality Factors Influencing Vegetable Prices". In: *Journal of Farm Economics* 10(2), pp. 185–196. ISSN: 10711031. URL: http://www.jstor.org/stable/1230278.