An approach for estimating real estate market dynamics

Daniela Martínez Duarte¹ and Jorge Finke²

Abstract—This thesis evaluates the performance of four machine learning algorithms to estimate the price of residential properties in the City of Santiago de Cali. Price estimation relies on measures of urban variables of street connectivity and techniques from information discovery. In particular, using the technique of maximal information coefficient (MIC), we demonstrate that attributes like the proximity to universities or shopping malls relates to housing prices to the same extent as the socio-economic strata. Furthermore, we show that there exist relationships between centrality measures of the street network, specifically, closeness and eccentricity, and housing prices. Finally, our approach enables potential buyers and sellers to determine which properties represent potentially good market offers. In general, it provides a data-driven understanding of the housing market and facilitates the process of decision-making for prospective buyers and sellers.

I. INTRODUCTION

The real estate market plays an important role in urban development. In general, variables like population growth, infrastructure development, and traffic are associated to price valuation and appreciation. Understanding the relationship between real estate prices and these variables is essential for urban planning and design.

Numerous approaches are being used to estimate residential property prices. The work in [1] presents perhaps the most common approach, in which the price of a house is determined by both internal (e.g., floor area, number of rooms, and age) and external characteristics (e.g., location, neighborhood amenities, accessibility, and environmental factors). This approach aims to identify the extent to which each characteristic influences housing prices. Although other studies follow similar techniques for estimating prices [2]-[5], there is no consensus on how to select the explanatory variables to be considered for estimation.

Some findings demonstrate that housing prices are closely related to location attributes as well as private and public service amenities [6]-[8]. The work in [6], for example, shows that schools, hospitals and metro stations influence the value of housing properties. For properties located in suburban area, accessibility to public transportation is a determinant factor. Other factors like school quality, transportation, population age and landscape amenities also influence housing prices [2, 5, 7]. Moreover, the work in [8] shows that the urban configuration of a city is critical in understanding the relationship between location externalities and property prices.

Recent information technologies play a key role in trying to understand the relationships that underlie real estate markets. They enable us to leverage data from a variety of platforms like Oxl, Ebay and Amazon, on which users buy, sell, and rent real estate properties. The amount of public information available on such platforms facilitates the development of new tools for capturing and analyzing large data sets, and for analyzing market dynamics. The aim of such analysis is to provide real-time insight that reveals hidden patterns of supply and demand, and to respond efficiently to emergent market forces. The quality and quantity of the online data sources becomes a key advantage to create knowledge about the market.

The motivation of this thesis is threefold. First, we want to find meaningful relationships that shape the housing market prices in the city of Santiago de Cali, using the maximal information coefficient (MIC) on various urban measures including measures based on the street network. Second, we aim to evaluate the performance of different machine learning algorithms in estimating the price of residential houses. Finally, we will design a web-based application that provides users a forecast of real estate.

II. LITERATURE REVIEW

A. The Street Network

There are two approaches to construct the street network a city. The first approach (known as the primal approach) is based on a planar graph in which nodes represent intersections and edges represent street segments. The second approach (known as the dual approach) considers street intersections as edges and street segments as nodes. Both approaches are widely used to try to understand the backbone or structure underlying the dynamics of cities [9, 10, 11].

B. Topological Measures

In graph theory, centrality measures determine the most important nodes in a network based on their relative location and interaction with other nodes. Commonly used centrality measures include betweenness, closeness, degree, eccentricity, and eigenvector [12].

Definition 1. The betweenness centrality of node \( x \), in a normalized form, is defined as

\[
C_b(x) = \frac{2}{(n-1)(n-2)} \sum_{i \neq j \neq x} \frac{\sigma_{i,j}(x)}{\sigma_{i,j}}
\]

where \( \sigma_{i,j} \) is the total number of shortest paths between nodes \( i \) and \( j \), \( \sigma_{i,j}(x) \) is the number of those shortest paths going through a node \( x \) and \( n \) is the total number of nodes in the graph. The measure of betweenness represents the
importance of a node in the flow of information of a network.

**Definition 2.** The closeness centrality of node $x$, in a normalized form, is given by

$$C_c(x) = \frac{(n - 1)}{\sum y d(x, y)}$$

where $d(x, y)$ is the distance between nodes $x$ and $y$, and $n$ is the total number of nodes in the graph. The closeness centrality indicates the ease with which a node reaches all others nodes in the network. Note that the nodes with a shortest distance to all others have a higher closeness centrality.

**Definition 3.** The degree centrality of node $x$ measures the number of edges connected to a node with respect to the total number of nodes without counting $x$. In other words

$$C_d(x) = \frac{\sum k(x)}{n - 1}$$

where $k(x)$ is the number of edges linked to node $x$ and $n$ the total number of nodes in the network. The degree centrality shows which nodes in a network have more connections. In a directed graph, there are two types of degree centrality, in-degree, which measures the in-coming connections to a node and the out-degree which measures the out-coming connections from a node.

**Definition 4.** The eccentricity of a node $x$ is the maximum distance from a node $x$ to any other node $y$ in a graph $G$, that is

$$C_e(x) = \max d(x, y)$$

where $d(x, y)$ is the distance between nodes $x$ and $y$.

**Definition 5.** The eigenvector centrality of node $x$ measures the importance of a node, based on the connections to other influential nodes in the network. The definition of this measure is

$$C_v(x) = \frac{1}{\lambda} \sum_{y \in V(x)} x$$

where $y \in V(x)$ correspond to the nodes connected to $x$ and $\lambda$ is a constant.

**C. Information Discovery Technique**

Consider a set of data $D$ composed by pairs of variables $X$ and $Y$. The relationship between a two-variable data is captured using rectangular grids on the scatter plot. These grids are divided in $r_x$ rows and $r_y$ columns to capture existing patterns in the data. Depending on the number of divisions of a grid, diverse non-uniform resolutions are evaluated to find the optimal data distribution. To determine this distribution, the technique of maximal information content (MIC) relies on the notion of mutual information (MI) to measure the statistical dependence between variables.

MI is defined through the probability distributions of a set of ordered pairs $(x, y) \in X \times Y$ as:

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)}$$

where $P(x, y)$ is the joint probability distribution function of $x$ and $y$, and $P(x)$ and $P(y)$ are their marginal probability distribution functions gencaga2014survey.

A common approach to calculate MI is using a fixed bin-width histogram-based method that calculates the probability of each bin, i.e., the number of points falling in a particular bin gencaga2014survey. MIC, adjusts the sizes of the bins to find the largest possible mutual information, which are denoted as $I^*(D, r_x, r_y)$. For each resolution, the grids associated to the optimal mutual information score are captured in a normalized matrix. The highest score in this matrix with the minimum number of rows and columns is selected as the MIC score reshef2013equitability. In particular,

$$MIC(D) = \max_{r_x, r_y < B(n)} \frac{I^*(D, r_x, r_y)}{\log_2 (\min(r_x, r_y))}$$

where $B(n)$ is a growing function that upper bounds the dimensions of the grids, depending on the sample size $n$ of the data $D$. It is defined as $B(n) = n^\alpha$. The work in [13] suggests to set the parameter $\alpha = 0.6$.

**D. Machine Learning Algorithms**

Machine Learning is widely used in a variety of applications. This sections briefly overviews the algorithms used in our work. The main concern of machine learning algorithms is to develop capacity of systematically learning from the data. There are two types of learning processes used in machine learning, named, supervised and unsupervised learning. Supervised learning requires a set of input variables to predict a (previously known) target variable. This type of learning can be used for classification or regression problems, depending on the type of the desired outcome. On the other hand, the unsupervised learning consists of finding meaningful patterns in data, without previously knowing the expected outcomes [14]. This thesis uses the supervised learning algorithms which are described below.

**Decision Trees:** A decision tree is an algorithms based on a tree structure, which consists in a set of nodes organized from top to bottom, starting from a root node. The tree is divided using decision rules into successive splits until the leaf or terminal nodes are reached [15]. Decision rules considered at each of the nodes depend on the value of features used to evaluate the conditions. These features can be categorical or numerical. In the case of numerical features, a threshold value is used as a reference to define
the resulting branches, which contain subsets of the initial dataset. The outcome of the model corresponds to the value at the leaf node where there are no further divisions.

**Random Forest:** Random forest is an algorithm used for both regression and classification problems. This algorithm first divides the training data set in random samples. It then uses a set of decision trees to estimate the value of the desired output. The decision trees, built randomly, are constructed using a subset of attributes to take appropriate decisions. The average of the output of the decision trees is considered the result of the random forest [14, p. 150].

**Gradient Boosting:** As for random forest, gradient boosting is also an algorithm implemented for both regression and classification purposes. This algorithm uses a combination of weak predictors to create a a more powerful predictor. Generally, the weak predictor models correspond to decision trees. The main idea of gradient boosting is to find successive models that reduce the error of the previous predictor and to combine them to minimize the overall error. The error in each step is measured using a differentiable loss function, which quantifies how well the model fits the expected result [16]. Selecting the proper loss function selection generally is application dependent.

### III. Results

#### A. Street Network

The network model of Cali is analyzed using both the primal and dual representations. To construct the primal representation of the street layout, we first retrieve the city of Cali as a polygon from Open Street Maps and then use the OSMnx library in Python to construct the network of public streets.

This network is represented as a weighted directed graph, in which intersections are turned into nodes and street segments into edges. The weight of each edge corresponds to the length in meters of the street segment. The edges are associated to the street names of the city and contained additional information on the type of the street (e.g. residential, secondary, tertiary or living street), the length of the street and its direction (i.e. whether is one-way or two-way). Conversely, the nodes are associated to intersections locations and to an unique ID.

Furthermore, to construct the dual representation of the street network of Cali, we convert the street segments of the city into nodes and associate a unique ID number to each of them. Moreover, we look at all the intersections of the street layout and determine which segments have common intersections. This street segments are linked with an edge using their ID as the reference.

Our approach for predicting the housing prices is based on the hypothesis that crucial information about this market can be acquired by studying the dynamics of a city with a model of the street network. To prove this hypothesis, we provide an statistical analysis relying on the maximal information coefficient (MIC) to find significant relationships between different variables associated to the urban network, including geometric and topological measures. MIC is used as the measure of correlation in this study. It identifies linear and non-linear relationships. Each segment is identified by means of its geographical coordinates, which correspond to the mid-point of each street.

The procedure for associating urban variables to the street network consists in taking measures from the mid point of each street. We consider various urban variables including the socio-economic stratum, the proximity to places like hospitals, bus stations, supermarkets, schools, churches, universities and shopping malls. Also, the centrality measures taken in the street network including betweenness, closeness, degree, eccentricity and eigenvector. Most importantly, the mean price of houses near the mid point of each street segment.

Once the urban variables and the mean price per square meter at different radii are associated to the street segments, we carry out an statistical analysis to find out if there is any correlation between them. To conduct this analysis, we compute MIC between pairs of variables \((x,y)\). The variable \(x\) represents the mean price per square meter of a specified radius and variable \(y\) represents the urban variables, including the social stratum , the proximity to different places, or the topological measures taken from the street network representations. The resulting MIC values, taking into account values greater than 0.4 indicate that the variables showing a significant relationship with the housing prices are the proximity to universities or shopping malls, the social stratum and the topological measures of closeness and eccentricity in the dual representation of the street network.

From the two street network representations, we choose the dual approach since the MIC values for two of the topological measures (closeness and eccentricity) considered in the statistical analysis have a stronger relationship with the mean price for a specific radius. Furthermore, street segments are the basis of the study conducted to understand the dynamics of the housing prices, which correspond to the nodes in this type of network.

The topological measures of closeness and eccentricity are closely related to the concept of accessibility. We infer from these results that the price of a house is influenced directly by its location, in terms of the access that it provides to all other points of interest in the city.

#### B. Model

Our approach proposes a model to estimate the price per square meter of a house using internal attributes of the house, as well as external attributes of the street network.
The construction of this model is divided in two parts. The first part consists of the development of a model to estimate the average price per square meter around a house. The input variables used for this model include the socio economic stratum, the proximity to a university, the proximity to a shopping mall, the closeness and eccentricity of the closest street segments to the house.

The second part receives the output result acquired from the previous model and internal attributes of the house corresponding to the floor area, the number of bathrooms and bedrooms are considered as inputs to the final model. To estimate the model that best fits our application we use different regression algorithms of machine learning. We adjust the parameters for each of the models to optimize their performance and use cross-validation to evaluate our final model of prediction.

Different statistical metrics are used to compare the performance of the tree machine learning algorithms of the first model. These metrics include the mean absolute error (MAE) and the $R^2$ (Coefficient of determination). The mean absolute error indicates the difference between the observed and predicted values, while the coefficient of determination indicates how well the model fits the observed values.

Figs. 1 and 2 show the MAE and the $R^2$ calculated in 10-Fold cross validation for the different machine learning algorithms. Figure 1 show that the MAE of gradient boosting model is lower compared to the random forest and decision tree models. Likewise, figure 2 demonstrates that the model that best fits the observed values based on the $R^2$ is the gradient boosting with a mean value of 0.95.

However, the difference between the mean $R^2$ of random forest and gradient boosting is less than 0.1. Another important conclusion from the previous results is that the variance of both the error and the performance values for the different folds is lower for the gradient boosting method. From these observations, we conclude that gradient boosting model is the preferred algorithm to predict the average price around a house.

For the house price estimation, we use data from OLX advertisements to train and test the model of prediction. 70% of the data is used for training the model and the remaining data is used for validation. We use the location of houses in the validation set to estimate the external attributes of the 3 closest street segments, then we calculate their mean values and run the previous model. With these results and the internal attributes of the house we compare the above machine learning algorithms. Figs. 3 and 4 show their performance.

The final prediction using gradient boosting has approximately 71% of accuracy, reaching a satisfactory performance. Furthermore, the mean absolute error is approximately of $340,000$ COP. This results suggest that our model estimates a higher price for some of the houses in the validation set as shown in figure 5 with the red points. On the other hand, results show that there are also a great amount of houses whose price according to the model should be lower (blue points in figure 5). This scenario allows us to analyze the dynamics of the housing market during the observation period.
From this analysis, we can conclude that the houses above the black dashed line (fit line) that have larger errors are opportunities of investment. The explanation of this result is closely related to the attributes studied by the model, because it means that based on the external and internal characteristics of the house, it should be more expensive. On the other side, the values whose price is over the estimation might be a consequence of the interaction of supply and demand in the market. In case of house-purchase the seller can assign a higher value to the house in order to negotiate its reduction when a client is interested in buying. Notwithstanding this case, is just one of many possible situations occurring in the housing market.

**Data:** The real estate data used for this approach is taken from OLX advertisement’s data set. The data gathered includes all residential properties located in Cali that were published from April to November 2018. The data set provides listings of properties with information on type, number of rooms, age, floor area, location, stratum and price. The process of preparation consists in cleaning, filtering and organizing the information. We remove the advertisements with missing data or entries with locations, which does not correspond to the neighborhood indicated in the description. Likewise, advertisements with properties whose number of bathrooms, number of rooms and floor area are unrealistic or filled with invalid values. In the case of price per square meter, the univariate outlier detection approach is used to identify outliers.

The univariate outlier detection approach uses the interquartile range (IQR) to find the existing outliers (i.e. the values that deviate from the majority of samples in the dataset). The latter is understood as the difference between the first and third quartiles (denoted as Q1 and Q3). The first and third quartiles correspond to the 25th and 75th percentile of the data. In particular, if the price is lower than \(Q1 - 1.5 \times \text{IQR}\) or greater than \(Q3 + 1.5 \times \text{IQR}\), the value is considered an outlier. The box plot in figure 6(a) shows the distribution of the price per square meter according to the stratum published and the histogram in figure 6(b) shows the number of data points per stratum available in the data set. Note that the highest dispersion in price occurs for stratum 1, which has a very limited amount of data. Other characteristics of properties and descriptive statistics are exposed in table I. The resulting sample contains 6003 properties.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floor Area (m²)</td>
<td>139.28</td>
<td>120.31</td>
</tr>
<tr>
<td>Social Stratification</td>
<td>4.28</td>
<td>1.15</td>
</tr>
<tr>
<td>Rooms</td>
<td>3.30</td>
<td>0.92</td>
</tr>
<tr>
<td>Bathrooms</td>
<td>2.68</td>
<td>1.10</td>
</tr>
<tr>
<td>Age of House (Years)</td>
<td>14.2</td>
<td>12.06</td>
</tr>
<tr>
<td>Price per Square Meter($)</td>
<td>2,412,939</td>
<td>880,304.4</td>
</tr>
</tbody>
</table>

**Fig. 6:** Descriptive statistics: (a) Boxplot of price per square meter for each stratum (b) Distribution of the stratum published in the resulting data set.

**C. Web Application**

The proposed approach provides a web-based application that gives users an estimate about the price per square meter of a property. The application is developed using a web framework written in Python known as Flask. The main purpose of the application is to design an interactive interface where users can easily receive feedback from the final ML model used for price estimation.

The development of the web application is divided in two main parts, the front end and the back end. The first part is focused on the design of the user interface and is made using HTML, CSS and JavaScript. Considering that
one of the most important parameters for our model is the property’s location, we use the mapping platform Mapbox to include an online map into our application. In particular, we use Mapbox.js that is a JavaScript library integrated with leaflet for creating an interactive map. Additional features that are included with leaflet are the search control that allow users to enter the respective address of their property and the marker to indicate the resulting location. For the search control, we use the Esri Leaflet Geocoder that interacts with the ArcGIS Online services to translate the address into geographic coordinates. Furthermore, we create a form in HTML to request the user information about the property.

The second part is developed in Python and involves all the background processes that provide functionality to the application. In this case, the back end uses the real estate properties database for training the final ML model, built using the scikit-learn library written in Python. As shown in the previous chapter, our approach uses two models to estimate the price. For this reason, we use a session in a flask to store the necessary information required in each of the models. The trained models are saved in .pkl files that include their fitted parameters, which are later used for price estimation. The interaction between the front end and back end parts of our application is performed using the HTTP method known as POST in URL routing.

IV. CONCLUSION AND FUTURE WORK

The proposed approach enhances our understanding of the dynamics of housing market in the city of Santiago de Cali. It evaluates different machine learning algorithms to estimate the price of a residential property. Our results show that gradient boosting offers the best model for price estimation. Key urban variables that explain house price are the floor area the property, the proximity to the socio-economic strata ranging from one of the most important parameters for our model is the property’s location, we use the mapping platform Mapbox to include an online map into our application. In particular, we use Mapbox.js that is a JavaScript library integrated with leaflet for creating an interactive map. Additional features that are included with leaflet are the search control that allow users to enter the respective address of their property and the marker to indicate the resulting location. For the search control, we use the Esri Leaflet Geocoder that interacts with the ArcGIS Online services to translate the address into geographic coordinates. Furthermore, we create a form in HTML to request the user information about the property.

The second part is developed in Python and involves all the background processes that provide functionality to the application. In this case, the back end uses the real estate properties database for training the final ML model, built using the scikit-learn library written in Python. As shown in the previous chapter, our approach uses two models to estimate the price. For this reason, we use a session in a flask to store the necessary information required in each of the models. The trained models are saved in .pkl files that include their fitted parameters, which are later used for price estimation. The interaction between the front end and back end parts of our application is performed using the HTTP method known as POST in URL routing.

IV. CONCLUSION AND FUTURE WORK

The proposed approach enhances our understanding of the dynamics of housing market in the city of Santiago de Cali. It evaluates different machine learning algorithms to estimate the price of a residential property. Our results show that gradient boosting offers the best model for price estimation. Key urban variables that explain house price are the floor area the property, the proximity to the closest university and to the closest shopping mall, and the topological measure of closeness.

Combining centrality measures from graph theory and geometric measures on the euclidean space, the proposed approach evaluates the relationships between urban variables and housing prices. Our analysis relies on the maximal information coefficient (MIC) to uncover hidden such hidden (linear or non-linear) relationships. Furthermore, the analysis shows that the dual approach to characterizing street networks offers a more suitable representation for associating measures of street segments to housing prices.

The proposed approach can be extend to other cities to identify variables that shape local housing markets. Here, we considered the data from OLX postings for Cali, within an observation period from April to November, 2018. Our analysis enables us to pinpoint properties with a higher than expected estimate, but more interesting to uncover various residential properties with price estimates lower than expected. The latter represent properties with a high potential for investments from which high profits may be earned.

Based on our experience, we must highlight some important limitations of our work. First, the data set used for this approach is unbalanced, since it has more properties in the socio-economic strata ranging from 3 to 6. Moreover, the amount of data points with invalid locations or wrong entries is relatively high. Our analysis required to filter data in great detail to obtain the valuable information. Finally, the method used to find the average price around a house can be improve by using different geometric shapes to establish boundaries between neighborhoods and different socio-economic stratum; and urban variables analyzed can be extended and adapted to the particularities of the city.

REFERENCES