NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Lisbon's Real Estate Analysis based on Proximity Calculations

Maria Madalena Jorge Do Nascimento Valério

Project Work submitted to International Journal of Scientific and Research Publications

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

LISBON'S REAL ESTATE ANLYSIS BASED ON PROXIMITY CALCULATIONS

by

Maria Madalena Jorge do Nascimento Valério

Project Work presented as partial requirement for obtaining the Master's degree in Advanced Analytics, with a Specialization in Business Analytics

Supervisor / Co Supervisor: Miguel Castro

Co Supervisor*:* Bruno Jardim

02 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, February 2023

ABSTRACT

The real estate market is always changing and evolving and with sustainability becoming an increasingly important topic, the way the price of a property is determined, and the factors taken in consideration, should evolve as well. Inspired by the popular concept of Smart Cities, more specifically the 15-minute city approach, which is a concept highly focused on accessibility and walkability in cities, different variables were calculated to assess each property's accessibility and diversity of amenities. Both these factors are different for each property, depending on their location. This work presents an analysis of the real estate market in Lisbon where, aside from the physical attributes of a habitation, the diversity and accessibility to difference services in each location will be evaluated and integrated in the machine learning process. The goal is to know the impact of each calculated measure when predicting the price per meter value of each house, in order to help understand why similar houses across Lisbon have such distinctive prices. Both the Euclidean Distance and the Network Distance were used in the calculations. The distance to Tejo River and the number of commercial establishments within a 15-minute walk radius were two of the most important features in the predictive models tested. Three different methods were tested and improved, electing the Random Forest Regressor as the best the one and the one to be used in the final model. The final model had half of the variance in the target explained by the all the calculated features, which makes this analysis a potential good tool to help fill the gap of a predictive model that only factors in the physical characteristics of a house.

KEYWORDS

Lisbon; Smart City; Real Estate; Accessibility; Mobility

Sustainable Development Goals (SGD): This Master's Thesis contributes to the Sustainable Development Goal of Sustainable Cities and Communities.

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

MAPE Mean Absolute Percentage Error.

MSE Mean Squared Error.

R^2 R-Squared.

SUSHI Sustainable Historic City Districts.

VIF Variance Inflation Factor.

1. INTRODUCTION

Over the years cities have become vast areas, very often with car-oriented developments, characterized by a less appealing aesthetics of landscapes and an ever-growing population, so much so, that it is expected that hundreds of people will migrate to cities in the next decade (Luscher, 2021). Faced with this future scenario, the recent health crisis of Covid-19 and the prolongated climate breakdown, more cities are shifting their focus to becoming more sustainable, inclusive and with quicker ways of commuting (Pozoukidou & Chatziyiannaki, 2021). These crises ended up exposing the true problems and fragilities within cities and their need for a response. One of the most striking realizations was the time saved by working remotely Before the global Covid-19 pandemic people would spend an extensive period of time commuting to work every day, which not only is very time consuming but also very harmful for the environment and public health (Johansson, 2017). This shift from the traditional office work and rigid workspaces to different work styles, taking in consideration what people enjoy and need to do (Taylor, 2021), has helped 15-minute cities allocate workspaces to specific areas, allowing companies to have other small offices across town where their employees could go. Not only it promotes a better life-work balance, leaving more time for people to do what brings them joy, while not commuting to work every day, but also it helps to reduce emissions and diminishes the transferability of COVID-19 (Moreno, Allam, Chabaud, Gall, & Pratlong, 2021). According to an international coalition of mayors focused on climate change and sustainability called C40 Cities, implementing the idea of a 15-minute city could help the urban areas recover from the financial and economic devastation of the pandemic (C40 Cities Climate Leadership Group, 2021). As the 15-minute city concept grows stronger, neighborhoods within cities will be fully focused on fulfilling needs like accessibility, walkability, residence density and land mix use, restoring the urban concept of proximity. Making the most out of the close location between services, activities and enabling people to access different opportunities within their urban environment (SmartCity, 2022). This proximity-based strategies offers people local access to a wide scope of amenities, like schools and preschools, healthcare services, social services, restaurants, entertainment and cultural events, parks etc., everything that is crucial for quality of life (Boucher, 2020). This transformation, however, might be a double-edged sword. As neighborhoods progress into small walkable areas, the real estate value of the proprieties within those neighborhoods may very likely increase, given the fact that one of the most important aspects of a propriety is its location and its proximity to points of interest.

Lisbon has become one of Europe's smartest cities, having received the *European Green Capital Award* in 2020 (Smart City Lisbon, 2018) and, over the last decade, amplified the available ways of commuting in the city by building more bike paths, rearranging key locations like squares, roundabouts and streets. One of the biggest purposes of the strategic plan of Lisbon for 2020 was to increase its population by promoting housing, taking smart-city initiatives regarding daily life and ageing (POR Lisboa 2014-2020). In 2019, before the Covid-19 pandemic, Portugal had one of the most dynamic real estate markets within the Western Europe, mostly due to its tax incentives granted to foreign buyers and their gold visas (Almeida, 2019). In these last few years, Lisbon has become a tourism magnet with many foreign real estate investors that renovate properties into short-term lease properties like Airbnb (Warren & Almeida, 2020). These transformations have caused an increase on Portugal's real estate prices given that now they are directed to tourists, who have greater purchasing power. According to the Eurostat, back in 2019, Portugal had an increase of almost 10% in their real estate prices, the biggest within Europe (Idealista, 2021). As Lisbon and its neighborhoods continue to change and become more sustainable, inclusive and with amenities close proximity, it is natural to assume that the real estate prices in the capital will follow these improvements and, ultimately, increase.

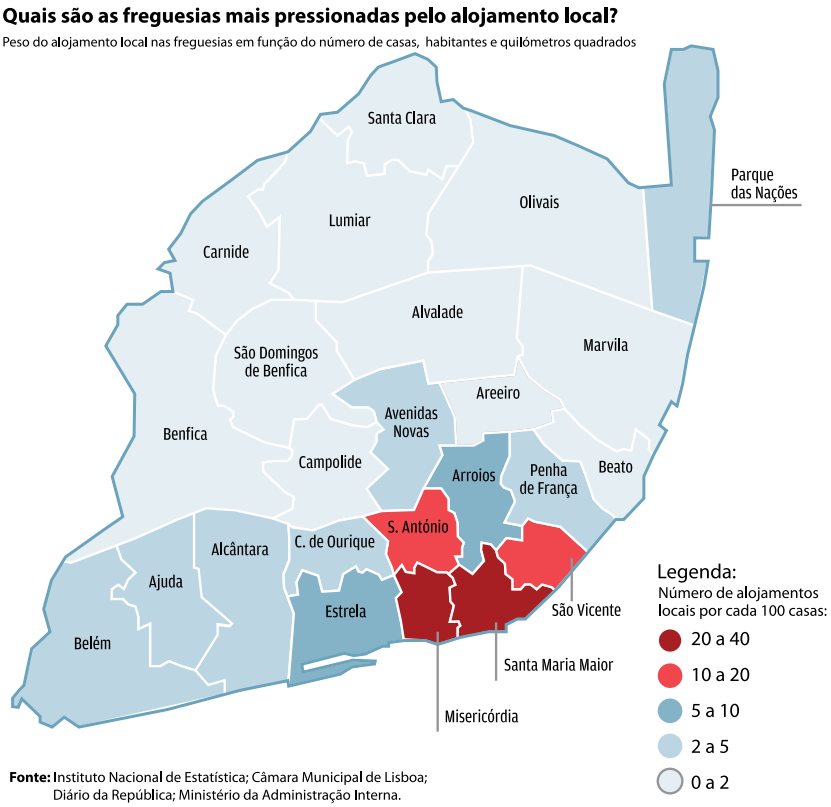


Figure 1.1 – Lisbon's parishes most pressured by local accommodation. Retrieved From https://tinyurl.com/jornal-de-negocios

Having said this, the predictive models created, as well as appraisers, might not be taking in considerations the accessibility aspects of each real estate location and therefore wrongly assess its value. To mitigate this small prediction errors and help understand why sometimes properties with the same characteristics have highly distinctive prices, the distance of each property to different amenities, as well as their count and the Shannon Diversity Index were calculated and used as input features in our models. Said models that used different predictive methods in order to understand which would be as the final model deployed.

For this analysis, Confidencial Imobiliário, a data bank with statistical data about transaction prices and residential lease agreements in Portugal, provided data from 14.299 habitations, containing their characteristics and coordinates. The findings indicate that there is indeed a connection between the accessibility and the diversity of amenities within a 15-minute walk radius of each location and their price per meter. The results also indicate that a property's distance to Tejo River and its number of commercial establishments within a 15-minute walk radius play an important role in the final predictive model deployed. With these results, it can be said that this investigation is a good tool to be used alongside more traditional predictive models that mostly focuses on the physical aspects of a real estate. This investigation has also shown some correlations between amenities, the presence of some amenities influences the presence of others.

The remain chapters of this documents are structured as follows: *Chapter 1. Literature Review, Chapter 2. Methodology, Chapter 3. Results and Discussion* and *Chapter 4. Conclusion*.

2. LITERATURE REVIEW

The COVID-19 pandemic had a great socio-economic impact on cities and their residents, especially when, to guarantee fair levels of health, they had to endure lockdowns (Taylor, 2021). All the safety measures taken during the pandemic ended up exposing the weaknesses of cities, whether it was a long response time for an ambulance to get to an emergency or that there are no supermarkets nearby. These aspects affect the buyer's decision when looking for a property, its location. For example, sometimes the most crowded and with most traffic neighborhoods are usually the most expensive ones, no wonder why "location, location, location" is a common mantra in real estate (Struyk, 2021). First, it's crucial to understand what a "smart city" is, along with the 15-minute city approach, introduced by Carlos Moreno back in 2016. Only then the real estate market and the aspects taken in consideration when evaluating a property, can be explained.

2.1 THE 15-MINUTE CITY APPROACH

The "15-Minute City" approach presents a different perspective of "chrono-urbanism" in which a resident shouldn't take more than 15 minutes to reach his destination, whether commuting by bicycle or walking. Moreno believes that cities should be designed or redesign so that within those 15-minute distance residents are able to live, "the essence of what constitutes the urban experience is to access work, housing, food, health, education, culture and leisure" (Moreno, 2021). To reach this goal, a city should have the previously stated services in the vicinity and not only in the city center. The concept has four guiding principles that might change given the author's perspective, Moreno's original principles are *ecology*, *proximity*, *solidarity* and *participation,* however the principles taken in consideration for the analysis are *density*, *proximity*, *diversity* and *digitalization,* proposed by Zaheer Allam and Carlos Moreno after observing the difficulties most cities went through during the widespread of COVID-19 cases and subsequent health measures taken to mitigate the spread (Taylor, 2021).

Proximity: This principle is both temporal and spatial, which means that within the 15-minute accessible area, a resident in a given neighborhood can access any basic service. Proximity is essential to reduce both the amount of time lost in commuting inside the city as well as the environmental and economic impact caused by this action, In Paris this principle allowed them to maximize the exploitation of the available resources and infrastructures in the city using, for example, the school's playground would be transformed into parks to be accessed after school's hours (Allam, Moreno, Chabaud, & Pratlong, 2021).

Density: In this concept, density is viewed in terms of people per kilometer square. This principle is crucial when planning for a sustainable city, by knowing the optimal number of residents that can occupy a given area, governments can then effectively plan the space so that all the essentials are accessible to the people living there. wording the optimal density ultimately allows to pursuit economic, social and environmental sustainability (Allam, Moreno, Chabaud, & Pratlong, 2021).

Diversity: In the context of this concept, diversity is a twofold principle, it refers to the need of having mixed-use neighborhoods in order to provide a healthy mix of residential, entertainment and commercial components and also diversify the culture and the people. Having a mixed-use neighborhood will reduce the distance between home and the office or any other services making them less automobile dependent. Promoting an accommodatable city to different cultures and people will henceforth promote social cohesion, along with helping to increase social capital, crucial for a city to function effectively (Allam, Moreno, Chabaud, & Pratlong, 2021).

Digitalization: Digitalization is a vital principle for the "15-Minute City" because it ensures the actualization of the other three principles. The achievement of factors like inclusivity and real-time delivery is related to the effective deployment of different technologies. Digitalization has been very effective in services such as online shopping, cashless payments and more. This kind of availability,

within the 15-minute radius, would decrease the need for commuting since some services would be delivered in the comfort of home or the office. During COVID-19, digitalization made it possible for people to communicate virtually and work from home especially when more restrict health measures were imposed. This principle is seen to be a transversal element and it is essential in ensuring the successful implementation of the other principles (Moreno, Allam, Chabaud, Gall, & Pratlong, 2021).

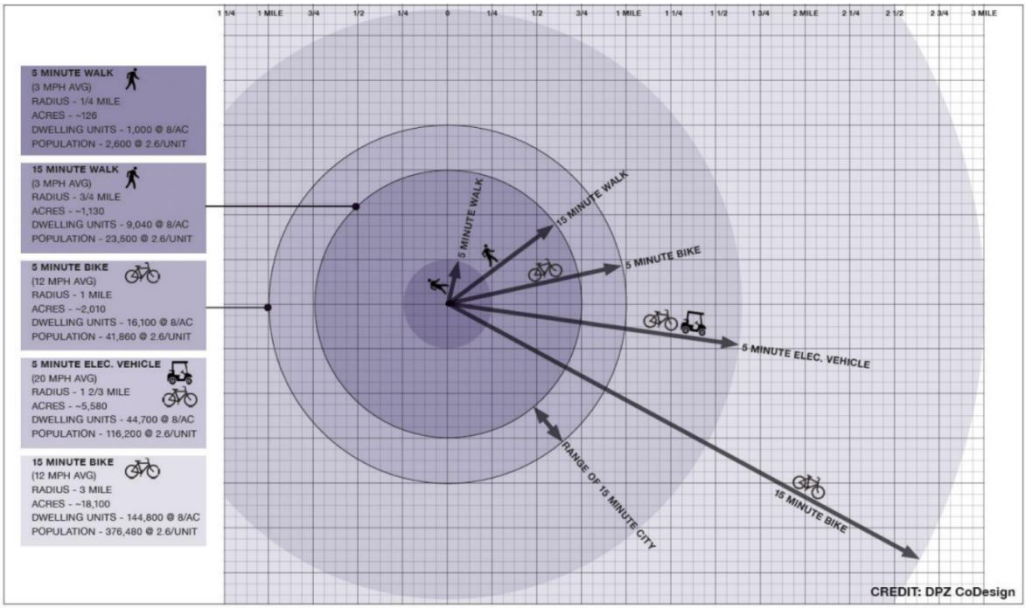


Figure 2.1 - Distance covered by each way of commuting. Retrieved From https://tinyurl.com/cnu-publicsquare

The key point in this concept is the proximity-based planning where a neighborhood in a city, accommodated with the optimal density, can access any basic service within a 15-minute walk or bike ride. The time saved commuting will give the residents the opportunity to interact with one another and participate in other social functions (Taylor, 2021). Exploring this micro-mobility will contribute to various aspects, in an economic, social and environment scale, by reducing congestion, pollution and then make the most out of the increased green spaces and planned structures. This concept has the potential to create new job opportunities, innovations and even reduce road maintenance and fuel costs (Duany & Steuteville, 2021).

City residents should have an agile, healthy, and pleasant life, not only for the most fortunate but for all social strata (Moreno, 2021). However, Philip Harrison, South African Research Chair in Spatial Analysis and City Planning, notes that as cities transform and are labelled as “smart”, there is a growing risk that it becomes too expensive for most residents to afford, due to expanded inequalities in the housing sector of the real estate market which appears to be growing at a disproportionate rate regarding the growth of the residence revenue (The Hindu Business Line, 2016).

2.2 REAL ESTATE

Whatever the real estate property in question, they are all characterized by a series of aspects that will either define them or group them with other products. These aspects help increase the products value and how interesting it is to the market. There are 6 aspects associated with real estate products: *(1) Location; (2) Dimension; (3) Function; (4) Age and Conservation Status; (5) Quality, Brand and Trend; (6) Aesthetics* (Costa, O produto imobiliário, 2018).

(1)- Location refers to the geographic context and the characteristics of the neighborhood in question. When it says that there is a property located in Portugal, Lisbon, in a dead-end street, with a school nearby and in which the houses there usually have only one or two rooms, it provides us with

two aspects that define the geographic location and the three aspects that characterize the surrounding area and its access, which are the first two and the last three respectively.

(2)- Dimension it's a fundamental aspect that characterizes our real estate product because it is often used as unit measure when assessing its value. It can be the number of square meters (or square feet, given the country we are in), the number rooms in a hotel or the number of divisions, which in Portugal is common to use "T2" when referring to a two-bedroom house with a living room.

(3)- Function or functionality is how good and adaptable a property is when being used as well as future problems that the property may have and its impact on the cost / functionality of it. For example, having a house with two bathrooms where one is a private bathroom connected to the bedroom and the other one is a small bathroom near the entrance of the house, is more functional than only having one bathroom for everyone that visits. Another example is living in an apartment where the building has an elevator, it is clearly more functional than having to come up the stairs.

(4)- Age and Conservation Status are also two important aspects to take into consideration. A very old building will certainly have historical, cultural and aesthetic features. That why a building's age should be just that, the number of years since construction. It is the conservation status that will tell us how attenuated those years are, whether it has "resisted well" to the effect of time. The real estate product is perishable which means it will lose quality over time and therefore loses value.

(5)- Quality, Brand and Trend are three aspects very close to each other, quality sometimes is associated with a specific brand or a current trend thus quality is a very subjective aspect. For an engineer a high-quality product is well executed, with no flaws that has a life-expectancy adjusted to the products purpose but for the general consumer, quality is related to the brand or what's trending. In the real estate context, quality is associated with who constructed it, who designed it (Brand) or what kind of materials were used for the workmanship (Trend).

(6)- Aesthetics is the most complicated aspect to analyze, because it is a subjective aspect that is present in every analysis, whichever the real estate product in question. When we appreciate a real estate product, we easily get the impression if that product is pretty, ugly or just normal. Aesthetics functions as a demand segmentation factor, to help people find what they are searching for. However, there are situations (usually when it has characteristics close to an art object) in which the perspective of a property's aesthetic quality is universal, and the real estate market recognizes its beauty or ugliness as universal.

When estimating a property value, besides all economic factors that will affect its value, one of the fundamental methods used is the *market method* (Costa, O Método de Mercado, 2018), which estimates the market value in the current day, for sale purposes. The appraiser will start by collecting all the elements available for the appraisal, plans, records etc. After collecting these information's, the appraiser will visit the property and do an inspection of the interior to check if the dimensions in the plans are correct, given that usually these plans don't have a graphic scale, only numerical, so the printing process might distort the true dimensions. Besides this, the appraiser will also inspect the state of conservation and occurrences of situations like damaged pipes or water infiltrations.

2.3 Evaluating Real estate's prices through 15-minute City features

As the 15-minute city concept becomes more popular and starts to be implemented in cities, it is important to define certain aspects to evaluate the changes, as well as their context. The following evaluation framework was built under three evaluation pillars: inclusion, safety and health. Presented by Pozoukidou, this framework with three key concepts represents "an alternative evaluation context that goes beyond the platitude of the general framework of urban resilience and sustainability of cities

and focuses on essential attributes that constitute and strengthen the concept of neighborhood as a place" (Pozoukidou & Chatziyiannaki, 2021). The proposed evaluation framework is illustrated bellow:

Pillars	Spatial Planning	Evaluation Attributes
Inclusion	Physical Planning	Housing: Variety and affordability of housing options
		Proximity to services: variety of services at place of residence
		Proximity to workplace: average time consumed to commute to work or distance to work from home
		Building density: average building density
		Land use mix: variety of land uses, including housing
		Accessibility: Access to rapid transit systems (rail, metro, tram)
		Multimodality: Alternative modes of transportation and their interconnections
	Community Building & Planning Process	Co-design processes for the production of space
		Bottom-up initiatives for the improvement of quality of life
Health	Physical Planning	Proximity to healthy and affordable food through fresh food markets and community urban gardens
		Proximity to basic health care
		Connectivity and multifunctionality of green and open spaces
		Active mobility (walking, biking, scootering etc.)
		Proximity to Cultural and Recreational opportunities
	Community Building & Planning Process	Cooperation of stakeholders and community for the interest of special groups (children, old people, people with disabilities etc.)
		Interaction between citizens in creating cultural, and recreational activities (urban gardening, walking teams etc.)
Safety	Physical Planning	Urban features that enhance the feeling of security
		Safe sharing of public space (including road space) for cultural and recreational activities
		Social distancing provisions due to COVID-19 restrictions
		Enhancement of safe mobility options due to COVID-19 i.e., road sharing practices
	Community Building & Planning Process	Lively neighborhoods in terms of the variety of activities in public space
		Participatory practices that include people of all age and abilities to combat physical and social isolation
Overall Proximity of Urban Amenities		Key resources localized in the neighborhood scale, including workplaces.

Figure 2.2 – Framework proposed by Pozoukidou & Chatziyiannaki (2021)

When evaluating a city using this framework, to use it correctly, it will be necessary to assess the effectiveness of the spatial planning, which can either be "Weak", "Medium" or "Strong". If the strategy's general objective, as well as its implementation measures and actions, are both explicitly referenced then it is a "Strong" point.

As mentioned previously, in another section of this paper, Paris is one of the most popular cities in the 15-minute city context. Under the leadership of Mayor Anne Hidalgo since 2014, who believes in this concept, Paris has become less of a car-oriented city by prioritizing bike lanes and public transports. "Paris En Commun" is a strategy presented in Hidalgo's re-election campaign back in 2020 which visions a city of Paris with revived neighborhoods, with increased community involvement and with measures and actions against the climate change (Pisano, 2020). This strategy has gained a lot of attention with COVID-19 and its recovery strategy. At the height of the pandemic 60 km of temporary bike lanes were introduced all over the city which later were made permanent, alongside other

measures in order to reduce air pollution, such as reducing the number of cars parking places to regain that space, increasing electric vehicle charging points, increasing tree canopy and enhancing pedestrian mobility (O'Sullivan, 2020). The greatest challenge in the 15-minute concept is to provide inclusive and diverse housing. In Paris there has been a decrease in housing accessibility and affordability due to gentrification and real estate speculation. The table below summarizes the evaluation of the "Paris En Commun" strategy based on the three pillars.

Pillars	Spatial Planning	Evaluation Attributes	Weak	Medium	Strong
Inclusion	Physical Planning	Housing	+		
		Proximity to services			+
		Proximity to workplace		+	
		Building density			+
		Land use mix			+
		Accessibility		Non-Applicable	
		Multimodality			+
	Community Building & Planning Process	Co-design processes		+	
		Bottom-up initiatives for the improvement of quality of life			+
Health	Physical Planning	Proximity to healthy and affordable fresh food			+
		Proximity to basic health care			+
		Connectivity and multifunctionality of green and open spaces		+	
		Active mobility			+
		Proximity to cultural and recreational opportunities			+
	Community Building & Planning Process	Cooperation of stakeholders and community			+
		Interaction between citizens			+
Safety	Physical Planning	Urban features that enhance the feeling of security			+
		Safe sharing of public space			+
		Social distancing (COVID-19)			+
		Safe mobility (COVID-19)			+
	Community Building & Planning Process	Lively neighborhoods			+
		Participatory practices			+
Overall Proximity of Urban Amenities					+

Figure 2.3 – Paris Framework Result presented by Pozoukidou & Chatziyiannaki (2021)

After analyzing Paris's framework, it is perceivable that most of the strategies are well implemented and organized however, there is a weak point. Due to its success on the other pillar's strategies, the city of Paris is struggling with the housing sector, having low variety and affordability.

3. METHODOLOGY

As previously seen, the 15-minute city concept has four principles that assess different aspects of a city. This research focuses on the principle of proximity and its impact on real estates, through various predictive models constructed, based on calculated variables like the number of amenities in a 15-minute walk radius and both the Euclidean and Network distance between an amenity and a property, excluding all the information regarding the physical aspects of the latter. These calculated variables try to characterize and score each property based on the 15-minute city's principle of proximity. This was the selected principle for the analysis given it highly focuses on accessibility and mobility in a city.

3.1 DISTANCE CALCULATIONS

Both the Euclidean distance and the Network distance were used to calculate the distance between amenities and each property, both calculated in very distinct ways. The Euclidean distance calculates the distance between two points by using the Pythagoras theorem (Equation 1), as for the Network Distance, it uses Lisbon's pedestrian network to find and calculate the best route from one point to another, using nodes and a search radius. There is no specific formula for network distance as it depends on the topology and configuration of the network being used.

To be more precise, the methods behind these distance values are the following:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2} \quad (1)$$

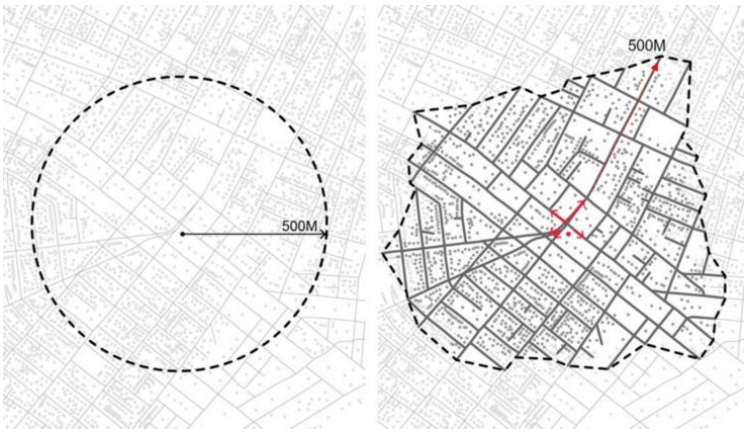


Figure 3.1 - Euclidean distance (left) Network distance (right)
Retrieved from - https://tinyurl.com/medium-AxU-platform

The scores for each property are divided in type of amenity, calculation method and scale, making it a total of 47 features. For example, there is a feature that calculates the network distance to emergency establishments and discards the 5 minutes of the trip, one feature that calculates the Euclidean distance to the water, which in this case is the Tejo River, and many others.

These calculated variables contemplate several scenarios to try to characterize the location of each property as best as possible. *Table* 1 contains the details and calculation methods for some of the features. The features that are missing are calculated the same way and they can be analyzed in Annex 1.

Name	Description	Calculation	Name	Description	Calculation
Transport Count	Number of Public Transportation option in a 15-minute walk radius	groupby(['IMOVEL_ID', 'transport_dist']).agg({'geometry': 'first', 'transport_count' : 'count'})	Commerce Count	Number of commercial establishments in a 15-minute walk radius	groupby(['IMOVEL_ID', 'comerce_dist']).agg({'geometry':'first', 'comerce_count':'count'})
Shannon Diversity Index	Measures the diversity of amenities in a 15-minute walk radius	$H' = -\sum_{i=1}^{S} pi\ ln\ pi$ **S** – Nr of the amenities **pi** – Nr of establishments by type of amenity	Average Network Distance	Average of the time of all the different amenities within a 15-minute walk radius	Avg Network Distance = distances.groupby("index_o", group_keys=False)['distance'].mean()
Transport Distance	Euclidean distance to the closest public transportation option	ckd_tree= cKDTree(B) dist,idx= ckd_tree.query(A, k = 1) idx = itemgetter(* idx) (B_ix)	Water Distance	Euclidean distance to the Tejo River	Water Distance = house_node.geometry .distance(river_node.geometry)
Government Score	Sum of the distances, in minutes, to all governmental buildings in a 15-minute walk radius	Score = sum([1/(1+dist) for dist in a['distance']])	Government Score 2.5	Sum of the distances to governmental buildings, discarding 2.5 minutes	Score_2.5 = sum([1/(1+dist+2.5) for dist in a['distance']])
Government Score 5	Sum of the distances to governmental buildings, discarding 5 minutes	Score_5 = sum([1/(1+dist+5) for dist in a['distance']])	Government Score 10	Sum of the distances to governmental buildings, discarding 10 minutes	Score_10 = sum([1/(1+dist+10) for dist in a['distance']])

Table 3.1 – The calculated features, its description and formula.

3.2 METHODS TO UNCOVER RELATIONSHIPS

With the remaining independent variables, different feature importance techniques were calculated to furthermore improve the predictive power of the final model. Each technique calculates a score for the input variables in that given model. The methods tested in each model were the Linear Regression, Decision Tree Regressor, and the Random Forest Regressor. Lasso and Spearman correlation were also analyzed.

3.2.1 Linear Regression

The linear regression uses the value of an independent variable to predict the value of the dependent variable, through a linear approach. This analysis allows the user to assess the quality of the variables in predicting the outcome along with which of these variables are more significant and how each of them impacts the dependent variable. In this phase of the research the Linear model created to test the feature importance is simply an instance, with no parameters associated. The linear regression formula represented below (Equation 2).

$$Y = \beta_0 + \beta_1 X \quad (2)$$

Where:

- $\mathbf{Y}$ – Dependent Variable
- $\boldsymbol{\beta_0}$– Constant / Intercept
- $\boldsymbol{\beta_1}$ – Slope / Coefficient
- $\mathbf{X}$ – Value of the independent variables

A common phenomenon in a regression analysis is the Multicollinearity, it exists when multiple independent variables are correlated in a model. The Variance Inflation Factor (VIF) identifies multicollinearity by estimating how much the variance of a regression coefficient is amplified due to multicollinearity in the model. VIF ranges from 1 upwards, where 1 is considered not correlated and values greater than 5 is considered highly correlated. The threshold used in this case was to retain the variable with a VIF value below or equal to 10.

3.2.2 Decision Tree Regressor

This method uses a structure flowchart-like where all the possible ranges of results are contemplated. In a decision tree there's nodes and branches, which contain the condition and the result, respectively. Each observation goes through this flowchart-like tree based on the value of its variables and a model is trained to predict future observations. Similarly to the Linear Regression, only an instance of the model was created for this part.

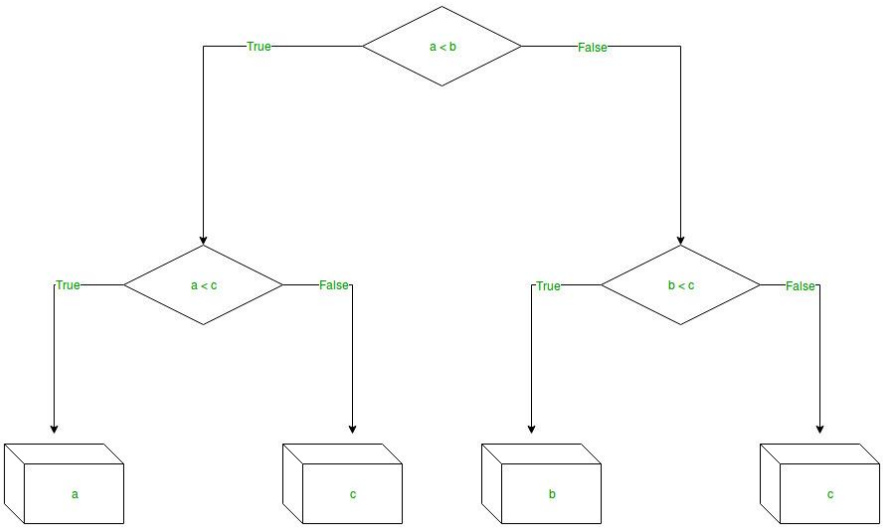


Figure 3.2 - Decision Tree Regressor. Retrieved from https://tinyurl.com/geeks-for-geeks

3.3.3 Random Forest Regressor

The Random Forest Regressor is an ensemble method composed of multiple decision trees. The final prediction for each observation is the average result of all the outcomes predicted by each tree. Generally, returns better results than Decision Tree Regressor method. Like the two previous Regression models, only an instance of the Random Forest Regressor is created to check the feature importance based on this method.

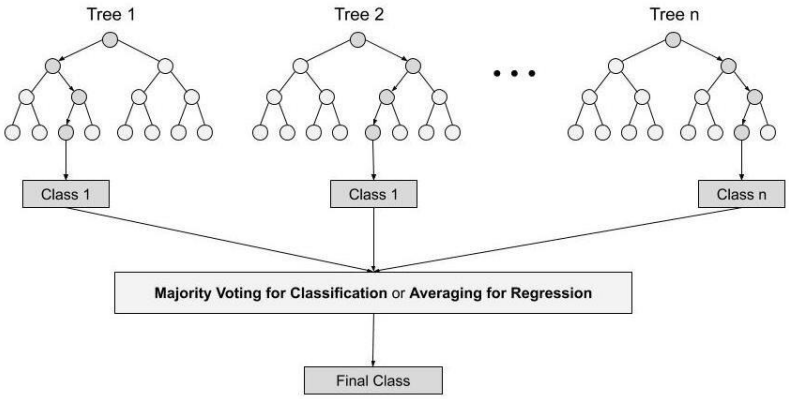


Figure 3.3 - Random Forest Regressor. Retrieved from https://tinyurl.com/analytics-vidhya

Besides this feature importance method two other feature selection techniques were tested, Lasso and the Spearman's correlation.

3.3.4 Lasso

The Least Absolute Shrinkage and Selection Operator is based on a regression analysis that focuses on enhancing the prediction accuracy of the models. This method is divided in two steps, the regularization of model parameters by shrinking the regression coefficients, and the feature selection, where every non-zero value is used in the predictive model.

3.3.5 Spearman correlation

Even though Spearman's correlation is not exactly a feature importance method, it is an important technique that measures the strength of association between two ranked features by assessing their monotonic relationships, whether linear or not. Spearman's correlation is, simply, the Pearson's correlation between the rank values of two features (Equation 3).

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{3}$$

ρ- Spearman's rank correlation coefficient

d_i- Difference between the two ranks of each observation

n – Number of Observations

6 – Constant Value

3.3 Evaluation Methods

To accesses each model's predictive power, R^2 Score, R^2 Adjusted Score, Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE) were calculated. These scores combined transmit all the information needed to understand if the selected variables are helping the models in predicting the target more accurately or not.

3.3.1 R^2 Score

R-Squared ranges between 0 and 1 which indicates the amount of variance in the output of the dependent variable that is predictable from the independent variable. A high value would mean that

most of the changeability in the dependent variable's output can be explained by the model. R-Squared indicates the proportion of points that lie within the line created by the regression equation (Equation 4), therefore, a higher value indicates better results.

$$R^2 = 1 - \frac{\sum_{1=1}^{n}(\hat{Y}_i - Y_i)^2}{\sum_{1=1}^{n}(Y_i - \bar{Y})^2} \quad (4)$$

3.3.2 R^2 Adjusted Score

As we add new features to a multiple regression model, R-squared will either continue to increase or stop, regardless of their quality. R-squared Adjusted eliminates this issue by taking in consideration the number of samples in the dataset, and the number of features in the model, increasing the score only when the newly added features improve the model's accuracy (Equation 5). Just like the R-Squared, a higher value indicated better results.

$$\text{Adjusted } R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1} \quad (5)$$

3.3.3 Mean Squared Error

The MSE (Mean Squared Error) measures the total error in a model by calculating the average squared difference between the observed values and the predicted ones. This calculation repeats itself for all the observations and afterward, all those values are summed and divided by the total number of observations (Equation 6). This method highlights the large errors, making it useful when working with models where these errors must be minimized.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (6)$$

In a regression, as the data points get closer to the regression line, the error in the model decreases, producing more precise predictions.

3.3.4 Mean Absolute Percentage Error

Mean Absolute Percentage Error, or MAPE, measures the accuracy of a predictive model by calculating the average percentage errors of the predictions. The error is the difference between the actual and forecasted value, so a smaller value would mean a smaller MAPE and therefore, better results.

3.4 SOFTWARE USED

To develop this study, Project Jupyter and python libraries were used. Jupyter notebook was the selected tool to develop and run the code for this analysis, due to its vast accessibility to python libraries. The OSMnx (Boeing, 2022) is one of the python libraries that was used, and it carried out one of the most important parts of the process which was collecting all the available data regarding all types of amenities (*government, health, education, finance, public transport, leisure*, *food* and *river)*. "OSM" means "Open Street Map" and it allows the user to download geospatial data, model and analyze real-world street networks. It allows users to model walkable urban networks with the use of python, where these models can be visualized and analyzed. The GeoPandas library was also very helpful in order to work with the coordinates that were given, and use them to calculate the difference distances, and therefore, their scores regarding the various amenities.

The following GitHub link contains two of the notebooks created to calculate the different features used in this thesis: GitHub - Thesis repository

4. RESULTS AND DISCUSSION

In this following chapter the results and findings of the analysis will be exposed and discussed, for a clearer interpretation of the outcomes, this chapter will be organized in four sub-chapters, 1.*Feature Analysis* 2. *Linear Regression Analysis*, 3. *Random Forest Regressor Analysis* and 4.*Discussion*

4.1 FEATURE ANALYSIS

According to the Lasso method, nine features were eliminated and thirty-five were selected, most of which also selected in the previous methods. Even though thirty-five were selected, only the best twenty features were taken into consideration. Some of the best features were the "Shannon Diversity Index" , "government score", and "education score 2.5". Lasso also allows to understand the impact of each feature, for example the feature "Shannon Diversity Index" has an importance value of -1423.94, which mean it is an important feature with an inverse correlation to the target. Some of the features selected by Lasso are represented below, the complete table is present in Annex 2.

Feature	Importance
Shannon Diversity Index	1423.94
Government Score	1033.45
Education Score 2.5	886.11
Government Score 2.5	847.97
Finance Score 2.5	526.85
Education Score 10	433.22
Health Score 2.5	418.10
Finance Score 10	331.98
Leisure Score 2.5	231.93
Food Score	207.02

Table 4.1 – Lasso's Top 10 Importance

The Spearman's correlation also offered some insights, not as concrete as the previous methods but nevertheless important. The most interesting insight was the correlation between the different scores which indicate that some establishments influence the presence of others. For example, the number of public transportations within a 15-minute walk radius of a house is highly correlated with the food establishments score near that same house, meaning that in areas where there are more restaurants, markets etc., the public transportation network offers more options for a person to commute. The same happens with the number of educational buildings, like high schools or universities, in a 15-minute walk radius and the distance to health establishments, meaning that schools often have heath and emergency services nearby.

Features that have a correlation higher than 0.8 should be discarded, at least one of them, since this means that both variables convey the same information. Having two very similar features, transmitting the same information, is not very effective.

The graphic below shows the Spearman's correlation between some of the variables.

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.
1.Commerce Count	**1**	0.7	0.6	0.9	0.8	-0.8	0.4	0.5	0.6	0.3	0.3	0.7	0.4	0.2	0.6	0.004
2.Public Transport Score	0.7	**1**	0.7	0.7	0.6	-0.5	0.6	0.6	0.8	0.2	0.02	0.8	0.3	0.5	0.8	0.06
3.Finance Score 10	0.6	0.7	**1**	0.7	0.5	-0.4	0.6	0.7	0.9	0.2	0.1	0.9	0.4	0.4	0.8	-0.2
4.Public Building Count	0.9	0.7	0.7	**1**	0.7	-0.6	0.5	0.6	0.7	0.4	0.3	0.7	0.5	0.3	0.6	-0.1
5.Transport Count	0.8	0.6	0.5	0.7	**1**	-0.5	0.3	0.4	0.5	0.3	0.03	0.6	0.4	0.1	0.5	-0.007
6.Shannon Diversity Index	-0.8	-0.5	-0.4	-0.6	-0.5	**1**	-0.1	-0.4	-0.4	0.08	-0.3	-0.5	0.06	0.008	-0.3	-0.008
7.Leisure Score 2.5	0.4	0.6	0.6	0.5	0.3	-0.1	**1**	0.6	0.6	0.1	0.08	-0.3	-0.5	0.06	0.008	-0.04
8.Entertainment Score	0.5	0.6	0.7	0.6	0.4	-0.4	0.6	**1**	0.7	-0.03	0.09	0.8	0.4	0.1	0.5	-0.3
9.Government Score 5	0.6	0.8	0.9	0.7	0.5	-0.4	0.6	0.7	**1**	0.1	0.2	0.8	0.4	0.3	0.8	-0.2
10.Education Count	0.3	0.2	0.2	0.4	0.3	0.08	0.1	-0.03	0.1	**1**	0.07	0.2	0.4	0.7	0.5	0.4
11.Average Network Distance	0.3	0.02	0.1	0.3	0.03	-0.3	0.08	0.09	0.2	0.07	**1**	0.02	0.2	-0.06	-0.03	0.2
12.Food Score 2.5	0.7	0.8	0.9	0.7	0.6	-0.5	0.6	0.8	0.8	0.2	0.02	**1**	0.3	0.4	0.8	-0.2
13.Greenspaces Count	0.4	0.3	0.4	0.5	0.4	0.06	0.7	0.4	0.4	0.4	0.2	0.3	**1**	0.2	0.4	0 .03
14.Education Score 10	0.2	0.5	0.4	0.3	0.1	0.008	0.3	0.1	0.3	0.7	-0.06	0.4	0.2	**1**	0.7	0.4
15.Health Score 2.5	0.6	0.8	0.8	0.6	0.5	-0.3	0.6	0.5	0.8	0.5	-0.03	0.8	0.4	0.7	**1**	0.1
16.Water Distance	0.004	0.06	-0.2	-0.1	-0.007	-0.008	-0.04	-0.3	-0.2	0.4	0.2	-0.2	0.03	0.4	0.1	**1**

Table 4.2 - Spearman's Correlation between some of the variables

Each amenity has a set of scores, but only one score of each set will be kept. Despite the fact that highly correlated features should be removed, in this case and since each variable refers a different amenity, some highly correlated features will be kept.

In the following sections other feature analysis techniques were deployed in order to better understand these calculated features.

4.2 LINEAR REGRESSION ANALYSIS

The Linear Regression representation between a target variable and an independent variable is one of the best ways to visualize their correlation and impact. Their coefficients transmit some of the most important information to understand their behavior, the constant and the slope. Table 4 contains the linear regression for each of the features selected for the final model.

Feature	β_0 / Intercept	β_1 / Slope	Feature	β_0 / Intercept	β_1 / Slope
Entertainment Score 5	3426.65	246.07	Health Score 2.5	3610.76	65.60
Leisure Score 2.5	3449.14	34.19	Greenspaces Count	3520.37	3.28
Shannon Diversity Index	5049.19	-1276.83	Public Buildings Count	3274.14	19.84
Public Transport Count	3261.35	7.88	Public Transport Score	3461.62	22.71
Education Count	3923.01	-8.44	Education Score 10	3933.27	-35.19
Food Score 2.5	3571.87	12.46	Govermment Score 5	3394.80	188.82
Commerce Count	3376.03	0.63	Water Distance	4519.41	-0.72
Finance Score 10	3404.08	106.01	Average Network Distance	3040.67	80.37

Table 4.3 – Linear Regression coefficients

With the values of β_0 and β_1 it is easy to understand each variable's behavior. The slope indicates if the feature has a direct or inverse of relationship through is sign, if the slope is negative then a lower value of the feature in question, will generate a higher Target value. For example, the relationship between the price per meter and the number of commerce establishments in a 15-minute walk radius, "commerce count", has a tendency of lower prices having a lower number of commerce establishments nearby. These variables have a positive linear relationship, and can be visualized below:

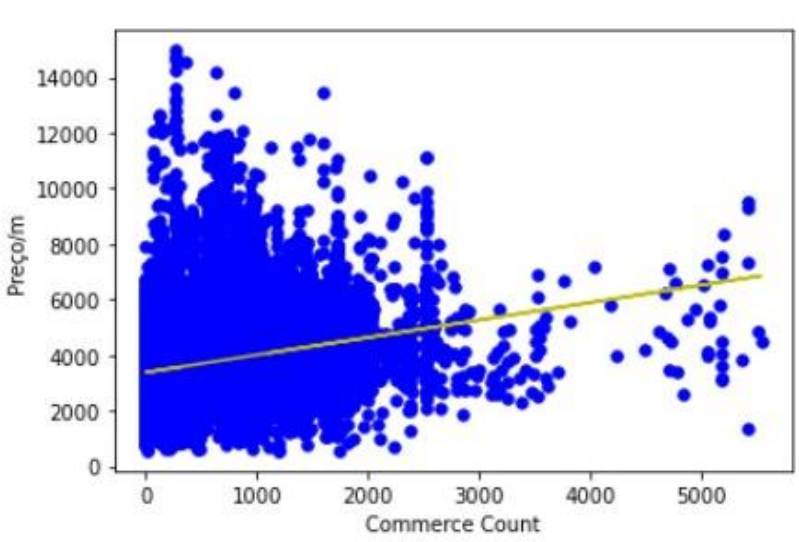


Figure 4.1 - Linear Regression between *Price/meter* and *Commerce Count*

When accessing the Linear Regression's feature importance results and excluding the variables with a VIF (Variance Inflation Factor) higher than ten, seven features remained which are also considered important in the different feature importance models tested. The selected variables by the Linear Regression's feature importance are the following:

Feature	VIF
Average Network Distance	1.7306
Water Distance	2.5302
Shannon Diversity Index	4.3115
Education Count	5.0916
Transportation Count	5.1829
Greenspaces Count	6.7085
Public Building Count	8.3501

Table 4.4 – Linear Regression features with a VIF below ten

The features above are inside the threshold, and there is no surprise in the outcome since all of these variables are recurrently labeled as important throughout the different feature important analysis.

Given that the Linear Regression is one of the simplest predictive methods, it is natural that the results are not optimal, however it is a very good method to gain some insights regarding each feature's correlation to the target.

Scores	Linear Regression
R^2 Score	0.1354
R^2 Adjusted Score	0.1344
MAPE	38.666
MSE	2659187.20

Table 4.5 – Linear Regression Model scores

With Linear Regression method, 13,54% of the variance in the target is explained by the all the calculated features and the average difference between the forecasted value and the actual value is 38,66%. For the mean squared error a value of 2 659 187,20 was obtained, which is very high and not a good result.

The Decision Tree Regressor is not displayed in the Results section since its results were similar to Linear Regression's and didn't offer any new insight.

4.3 Random Forest Regressor Analysis

Similarly to the Linear Regression, the feature importance and the model performance were analyzed. In some respects both methods had similar outcomes, particularly regarding the feature importance analysis. The Random Forest Regressor method considered the following features as it's the most important:

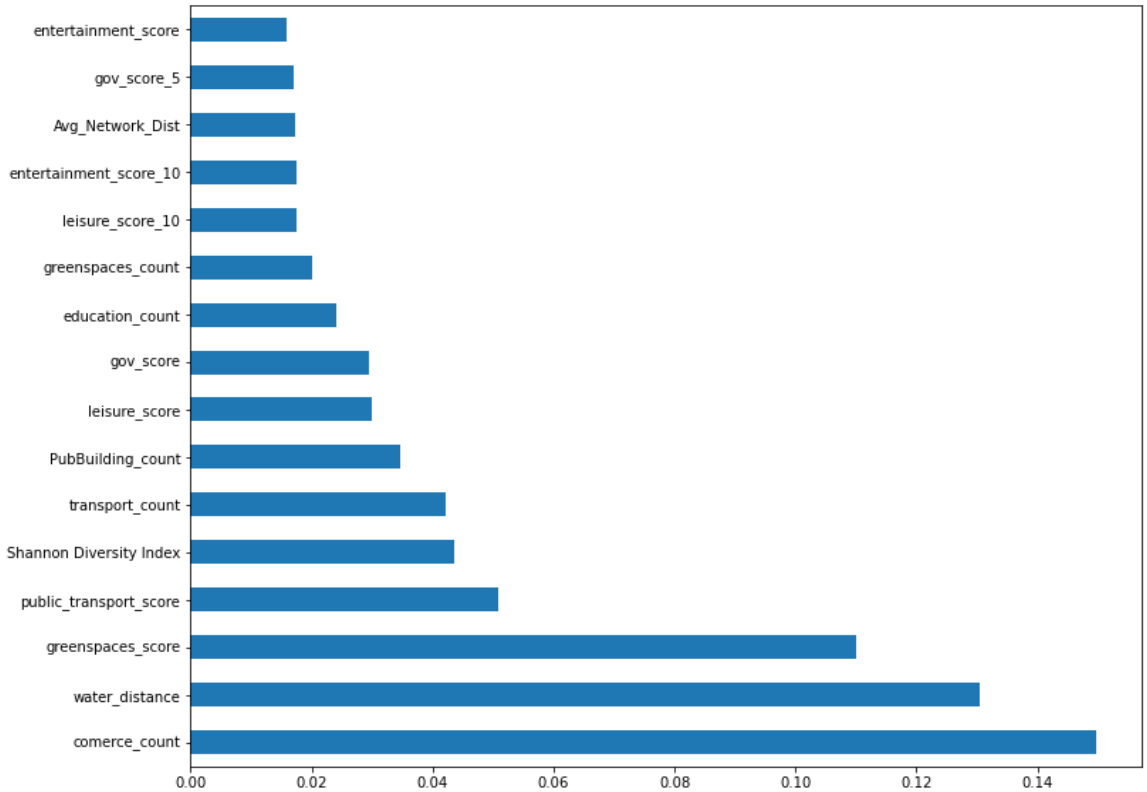


Figure 4.2 - Random Forest Regressor Feature Importance

When assessing the graphic above, certain groups of features are recurrently selected like the variables that count the establishments within a 15-minute walk radius, which are all present in the twenty most important features of all the different techniques deployed. The water distance variable is also a good example since is selected by all the feature importance methods.

The performance of this model was also tested, and its result are displayed below. The results are substantially better than the other models, even though the mean squared error value is still considerably high.

Scores	Random Forest Regressor
R^2 Score	0.4728
R^2 Adjusted Score	0.4730
MAPE	28.106
MSE	1729960.44

Table 4.6 – Random Forest Regressor scores

The results from the Random Forest Regressor model deployed can be considered quite satisfactory, since these features were calculated and they are not tangible characteristics of a property, the outcome could have been significantly worse. With 47,28% of the variance in the target is explained by the all the calculated features and the average difference between the forecasted value and the actual value is 28,11%.

4.4 RESULTS COMPARISON

The previous results of both the Random Forest and Linear Regression were simple instances of the methods, no parameters were searched in order to improve the scores. By creating only an instance the conditions are the same of each method and therefore, the results will be unbiased. When comparing the results between these two models, it is clear that the Random Forest Regressor outperforms the Linear Regression and, even though it is not represented in the results, it also performs better than Decision Tree Regressor, which is natural. As previously explained, the Random Forest is an ensemble of decision trees so it should perform better that a method that only uses one. Despite Linear Regression's performance results, this method is very useful to understand correlations between features and represent them, that is why this method is in the final part of this analysis. One of the findings were the variables considered important for both methods, which were mostly the same. "Water distance", "transportation count", "Shannon Diversity Index" and "Average Network Distance" were some of these variables, which ultimately confirms their importance when predicting the price per meter value of a house.

Regarding the models scores, not only a higher percentage of the variance in the target is explained by the selected features, the R^2 Score, but also MAPE and MSE scores are lower in the models that use the Random Forest Regressor, meaning the error are smaller. In spite a good evolution in the scores, some of the results are still considerably high, like the MAPE and MSE scores.

For the final model, the best combination of parameter values was tested in order to find even better scores. The Grid Search was not used since it was too time consuming, instead each parameter was

test and then added to the set of parameters. The model improved on all the other scores thus being considered the final model.

Scores	Final Random Forest Regressor
R^2 Score	0.4984
R^2 Adjusted Score	0.4979
MAPE	28.247
MSE	1542467.02

Table 4.7 – Final model scores

This would mean that, in the final model, almost 50% of the variance in the target variable is explained by the selected features, the predictions are, on average, 28,55% away from the targets. The results may not be optimal, but they can be considered satisfactory.

4.5 Discussion

The following section highlights the most important insights and conclusions that can be taken from this analysis, understanding which behaviors were expected and which weren't. Said analysis will start by the calculated features, where the most important features are as follows:

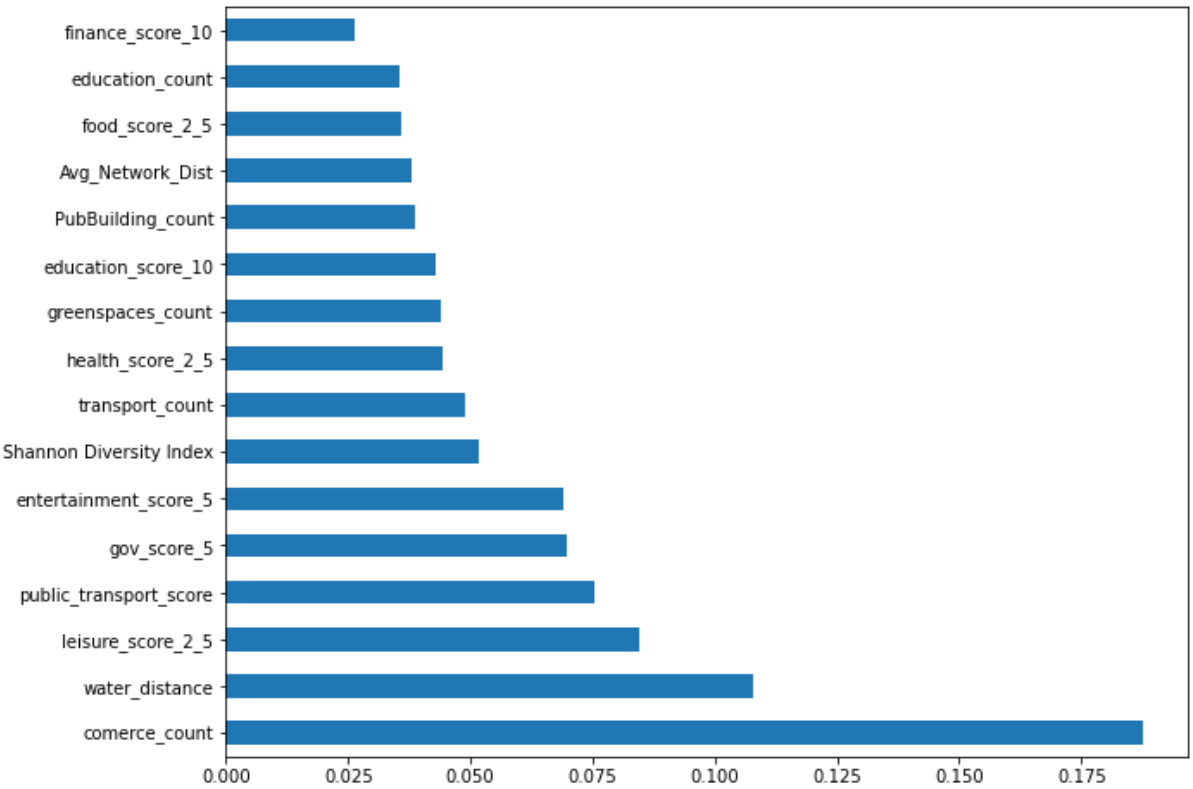


Figure 4.3 – Final model's most important features

Features like water distance, commerce count and public transport distance, were considered some of the most important features in the final model, which makes sense since most the times these are actual factors that may influence the final price of a house. The town center is, by definition, the main business and commercial area of the city thus, having a higher commerce count would imply that the property is in a more centered area of the town and therefore, more expensive. As it was represented previously, the commerce count has a positive linear regression, meaning that an area with a higher number of commercial establishments also has a higher price per meter.

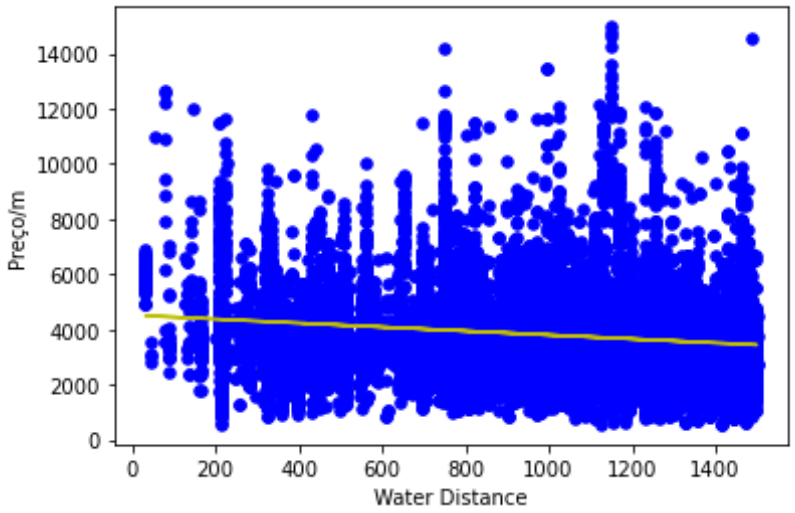


Figure 4.4 - Linear Regression between *Price/meter* and *Water distance.*

The figure above represents the relationship between the price per meter of a house and its distance to the water, which has a negative linear relationship. It is natural that houses closer to Tejo tend to be more expensive than the ones further away, given that people value more the natural landscapes. Usually houses that have an appealing view to the water, mountains or even to the city, tend to be more expensive. The price per meter of a house and its distance to the water has a negative linear relationship.

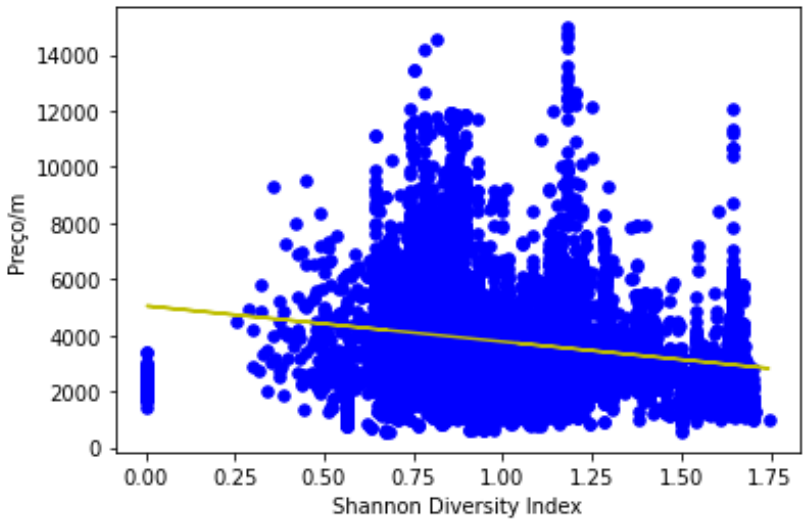


Figure 4.5 - Linear Regression between *Price/meter* and *Shannon Diversity Index.*

The Shannon Diversity Index, as represented above, has a negative slope which implies that a higher diversity level of establishments, within a 15-minute walk radius, have a lower price per meter value. This finding is very interesting, especially regarding the 15-minute city concept, given it shows that areas with a higher diversity of amenities are cheaper than the ones with less variety.

The results regarding the remaining features were as expected in most cases. The aspects, location wise, that people often consider important when looking for a house turned out to be the same ones that the final model finds more important, in most cases. Access to public transportation is one of them, people value the presence of public transportations, either to go to work or to easily commute inside the city without having to drive. One amenity score that could be considered important but didn't have a significant role in the predictions was the education score. Not only that but this score also had a negative linear relationship, meaning that a higher number of educational establishments and the closer a house is to one, the cheaper it will be.

To have an R-squared value of almost 50 % in our final model, having only calculated features inspired by the proximity pillar of the 15-minute city concept, confirms that accessibility factors do impact a house's final price. It may not be such a big impact like the physical characteristics of the property, but it definitely has a say in the final price and it should be considered future predictive models. In these features lies the answer to why certain houses, with similar characteristics, sometimes have very distinctive selling prices. In the end, it all about their location. With this being said, the predictive models created should have some of these accessibility features factored in, not only to mitigate the errors but also to better characterize a property.

With the 15-minute city concept growing, especially in the city of Lisbon, and one of its biggest concerns is the affordability and availability in the housing sector, like what happens in Paris. The next visualization contains some of the calculated features for two parishes in Lisbon, Avenidas Novas and Santa Clara, in blue and red, respectively. The values displayed in the chart for each feature are an average of all the observations for each zip-code. The complete plot of the radar graph, with a bigger range so that all the variables can be properly visualized, is in Annex 3.

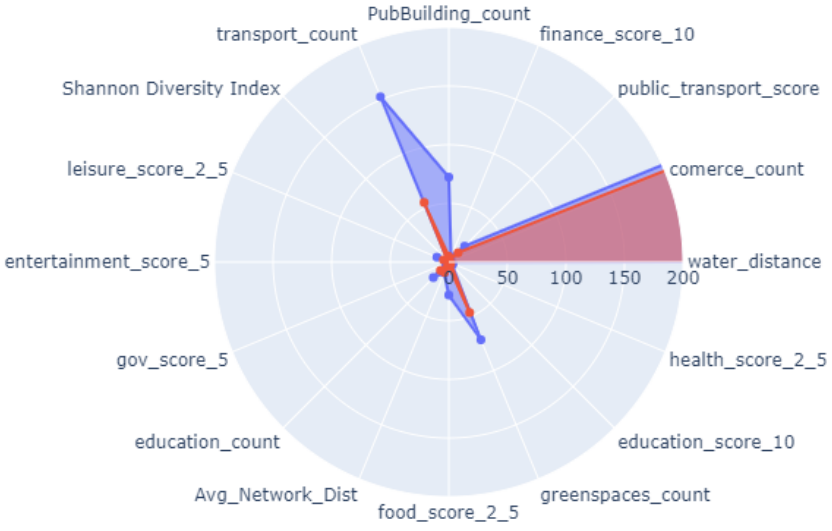


Figure 4.6 - Radar Chart of the features in Avenidas Novas (blue) and Santa Clara (red)

Avenidas Novas, the most expensive out of the two parishes, has better feature values. The number of public transportations is higher, the number of public buildings (ex: companies) is also higher and, as we can see in the table below, Table 9 ,the distance to Tejo river is significantly lower.

Features	Avenidas Novas	Santa Clara
Water Distance	1012.12	1302.51
Commerce Count	1652.74	278.144
Public Transport Score	19.11	11.208
Finance Score 10	4.90	1.216
Public Building Count	72.35	4.86
Transport Count	152.42	54.987
Shannon Diversity Index	0.83	1.020
Leisure Score 2.5	10.94	4.890
Entertainment Score 5	2.04	0.386
Government Score 5	2.93	0.596
Education Count	18.65	10.040
Average Network Distance	9.95	9.553
Food Score 2.5	28.11	3.063
Greenspaces Count	71.68	46.619
Education Score 10	3.92	2.265
Health Score 2.5	3.84	1.015

Table 4.8 – Feature value of Avenidas Novas and Santa Clara

Even though this analysis is not regarding the impact of the calculated features directly, it would be interesting to verify if indeed the most expensive areas in Lisbon are the ones with the best feature values. This last analysis should be applied to all the parishes of Lisbon and compared with its price per meter, a suggestion made in the *Future works* section.

5. CONCLUSION

Often, when appraising the value of a property, the level of accessibility and mobility is not factored in, the physical aspects such as the area and the number of rooms, are much more important when trying to predict the price per meter of a house, than the aspects regarding its location. However, it had to be understood the impact these other aspects could have on the final price. Inspired by the 15-minute city concept and its pillar of proximity, different variables were calculated in order to understand the accessibility around each property. A total of forty-seven features were created using Euclidean Distance, Network Distance, by counting the number of establishments by amenity and a few other methods. Not all the features should be factored in the final model thus, in order to select a set of variables, different feature importance and feature selection methods were analyzed. Lasso, Spearman's correlation and the feature importance of three regression methods were deployed and analyzed, leaving only sixteen features. These remaining features were used as input variables in three different models, the Decision Tree Regression, the Linear Regression and the Random Forest Regressor. To verify the quality of each model four different scores were calculated and, out of the three instances created, the Random Forest Regressor was the one that performed better. All scores had significant improvements, specially the R^2 score with a value of 0.5, meaning that half of the variance in the target variable is explained by the selected features. With the final model obtained, a deeper analysis into the most important features and their impact on the target was made. Calculated features such as "water distance", "commerce count", "public transport distance score" and "green spaces distance score", turned out to be quite important for all the models, proving that the accessibility and location of a house do have an impact on its selling price. When appraisers say that a property's location is one of the most important aspects, it is true. Even though finding and tunning the predictive model was not the main focus, the scores of the final mode were considerably good, especially considering that the inputted variables were idealized based on the concept of 15-minute cities, from which different scores, distances and counts of amenities were calculated.

In conclusion, this analysis alone will not predict the price per meter of a house as accurately as with its physical attributes factored in but, it will help explain the difference in pricing when two houses with identical physical characteristics have highly distinctive prices, as there is a correlation between these calculated attributes and the target variable that proves this theory.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

As stated previously, this analysis focuses on understanding if the proximity principle of the 15-minute city approach has any impact on the price per meter value of a houses. The variables that support this analysis were created based on what the principle assess rather than what might be best inputs for the predictions. Not only this but there may be more variables to characterize a property's location and accessibility, that were not considered in this research. The methods used for the calculations of said variables were highly time consuming, which made any alteration or correction, that forced a re-calculation, a considerable setback. Besides the technical limitations, there were also a few topics that were not discussed in this thesis, and it can be seen as limitations as well. Despite the confirmation that the calculated features impact the predictive models when estimating the price per meter, there is no analysis that confirms that the most expensive areas in Lisbon are in fact the ones that have a higher commerce count, or a lower water distance. Another limitation is that the impact of the calculated features is only tested for the city of Lisbon thus it cannot used as a fact went testing it in other cities. Lastly and probably the biggest limitation is that these predictions and analysis is based on the values of today and this is very volatile market, where the prices can easily change and vary from city to city.

For futures works, it would be advisable to develop a better predictive model, a deeper analysis regarding the machine learning part of this process, where a more complex predictive model, with the ideal parameters, would reduce the MAPE and MSE of our final model, as well as increase its R-squared and R^2 Adjusted score. It may be interesting to test this analysis in different cities, where the 15-minute concept is growing, and the real estate values might me shifting, in order to solidify this correlation between accessibility factors and the value of a property. This analysis might be a good tool to help fill the gap of a predictive model that only factors in the physical characteristics of a property. Regarding the limitations previously presented, there is a good opportunity for future works. A comparison between the price per meter of a houses, per parish and the values of the calculated features would confirm if indeed the most expensive areas have better accessibility factors or live closer to the river. To verify that this impact not only occurs in properties in the city of Lisbon, other cities can also be analyzed, like Paris, that follows the 15-minute city concept, or other cities that don't have their accessibility factors as well developed, and check if the same impact is observed or if it is different, and if it is extremely different, why does it happen.

REFERENCES

Allam, Z., Moreno, C., Chabaud, D., & Pratlong, F. (2021). *Proximity-Based Planning and the "15-Minute City": A Sustainable Model for the City of the Future.*

Almeida, H. (2019, September 19). *Europe's Hottest Property Market Is Getting Too Hot for Some*. Retrieved from Bloomberg: https://www.bloomberg.com/news/features/2019-09-19/portugal-is-europe-s-hottest-property-market-too-hot-for-some

Andres Duany, R. S. (2021, February 8). *Defining the 15-minute city*. Retrieved from CNU : https://www.cnu.org/publicsquare/2021/02/08/defining-15-minute-city

Boeing, G. (2022). *OSMnx 1.1.2. Retrieved from OSMnx: https://osmnx.readthedocs.io/en/stable/*

Borges, I. M. (2021). *Tourism and tranformation dynamics in historical neighborhoods: The case of Alfama.*

Boucher, D. (2020). *Local Living, Rise of 20 Minute Cities Post-Covid.* Retrieved from The Deck: https://thedeck.org.au/queensland-statewide-all-regions/local-living-rise-of-20-minute-cities-post-covid/

C40 Cities Climate Leadership Group, C. K. (2021, May). *Why every city can benefit from a '15-minute city' vision.* Retrieved from C40 Knowledge Hub: https://www.c40knowledgehub.org/s/article/Why-every-city-can-benefit-from-a-15-minute-city-vision?language=en_US

Costa, M. D. (2018). O Método de Mercado. In *O Valor do Imobiliário - Princípios, Fatores e Técnicas de Avaliação Imobiliária*.

Duany, A.; Steuteville, R. (2021). *Defining the 15-minute city.* Retrieved from CNU: https://www.cnu.org/publicsquare/2021/02/08/defining-15-minute-city

Gouveia, J. P. (2021, June). *City Transformation; Alfama - Lisboa (Portugal).* Retrieved from Sustainable Historic City Districts.

Idealista. (2021, January 21). *Property prices in Portugal have soared in the last decade*. Retrieved from Idealista News: https://www.idealista.pt/en/news/property-for-sale-in-portugal/2021/01/21/845-property-prices-in-portugal-have-soared-in-the-last-decade

Luscher, D. (2021, June 16). *Introducing the 15- Minute City Project*. Retrieved from :15 City: https://www.15minutecity.com/blog/hello

Moreno, C., Allam, Z., Chabaud, D., Gall, C., & Pratlong, F. (2021). Intoducing the "15-Minute City": Sustainability, Resilience and Place Identity in Future Post-Pandemic Cities. *Smart Cities*.

O'Sullivan, F. (2020) *Paris Mayor: It's Time for a '15-Minute City'.* Retrieved from Bloomberg: https://www.bloomberg.com/news/articles/2020-02-18/paris-mayor-pledges-a-greener-15-minute-city

Pisano, C. (2020). *Strategies for Post- COVID Cities: An Insight to Paris En Commun and Milano 2020. Sustainability.*

POR Lisboa 2014-2020. (n.d.). Retrieved from Lisboa Portugal2020: https://lisboa.portugal2020.pt/np4/%7B$clientServletPath%7D/?newsId=253&fileName=ISE_CH_FEDER_PI_9.7_9.8_10.5.pdf

Pozoukidou, G., & Chatziyiannaki, Z. (2021). 15-Minute City: Decomposing the New Urban Planning's Eutopia. *Sustainability*, 5 - 6.

Smart City Tech. (2022). HOW SMART CITIES COULD HELP US ACHIEVE EQUITY AND ACCESSIBILITY. Retrieved from SmartCity Press: https://smartcity.press/smart-cities-equity-and-accessibility/

Smart City Lisbon. (2018, September 03). Retrieved from Smart City: https://smartcity.brussels/news-598-smart-city-lisbon

Struyk, T. (2021, February). *The Factors of a 'Good' Location.* Retrieved from Investopedia: https://www.investopedia.com/financial-edge/0410/the-5-factors-of-a-good-location.aspx

Taylor, P. (2021, February). *Where Did Our Commute Time Go?* Retrieved from Paul Taylor: https://paulitaylor.com/2021/02/05/where-did-our-commute-time-go/

The Hindu Business Line (15 September 2016). Retrieved from: https://www.thehindubusinessline.com/economy/smart-cities-could-result-in-social-inequality-say-experts/article9111629.ece

TED (Director). (n.d.). *The 15-minute city | Carlos Moreno* [Motion Picture]. Retrieved from https://www.youtube.com/watch?v=TQ2f4sJVXAI

Warren, H.; Almeida, H. (2020). *Airbnb Hosts Resist Lisbon's Plan to Free Up Housing.* Retrieved from Bloomberg: https://www.bloomberg.com/graphics/2020-airbnb-short-let-reforms-lisbon/

APPENDIX

Name	Description	Calculation	Name	Description	Calculation
Transport Count	Number of Public Transportation option in a 15-minute walk radius	groupby(['IMOVEL_ID', 'transport_dist']).agg({'geometry': 'first', 'transport_count' : 'count'})	Commerce Count	Number of commercial establishments in a 15-minute walk radius	groupby(['IMOVEL_ID', 'comerce_dist']).agg({'geometry':'first', 'comerce_count':'count'})
Education Count	Number of Educational establishments in a 15-minute walk radius	groupby(['IMOVEL_ID', 'education_dist']).agg({'geometry':'first', 'education_count':'count'})	Public Buildings Count	Number of Public Transportation option in a 15-minute walk radius	groupby(['IMOVEL_ID', 'PubBuilding_dist']).agg({'geometry':'first', 'PubBuilding_count':'count'})
Emergency Count	Number of Emergency establishments in a 15-minute walk radius	groupby(['IMOVEL_ID', 'emergency_dist']).agg({'geometry':'first', 'emergency_count':'count'})	Greenspaces Count	Number of Public Transportation option in a 15-minute walk radius	groupby(['IMOVEL_ID', 'greenspaces_dist']).agg({'geometry':'first', 'greenspaces_count':'count'})
Shannon Diversity Index	Measures the diversity of amenities in a 15-minute walk radius	$H' = -\sum_{i=1}^{S} pi \ln pi$ **S** – Nr of the amenities **pi** – Nr of establishments by type of amenity	Average Network Distance	Average of the time of all the different amenities within a 15-minute walk radius	Avg Network Distance = distances.groupby("index_o", group_keys=False)['distance'].mean()
Transport Distance	Euclidean distance to the closest public transportation option	ckd_tree= cKDTree(B) dist,idx= ckd_tree.query(A, k = 1) idx = itemgetter(* idx) (B_ix)	Government Score	Sum of the distances, in minutes, to all governmental buildings in a 15-minute walk radius	Score = sum([1/(1+dist) for dist in a['distance']])
Education Distance	Euclidean distance to the closest educational establishment	Calculated the same way as the **Transport Distance**.	Government Score 2.5	Sum of the distances to governmental buildings, discarding 2.5 minutes	Score_2.5 = sum([1/(1+dist+2.5) for dist in a['distance']])
Emergency Distance	Euclidean distance to the closest emergency building	Calculated the same way as the **Transport Distance**.	Government Score 5	Sum of the distances to governmental buildings, discarding 5 minutes	Score_5 = sum([1/(1+dist+5) for dist in a['distance']])
Greenspaces Distance	Euclidean distance to the	Calculated the same way as the **Transport Distance**.	Government Score 10	Sum of the distances to governmental buildings,	Score_10 = sum([1/(1+dist+10) for dist in a['distance']])

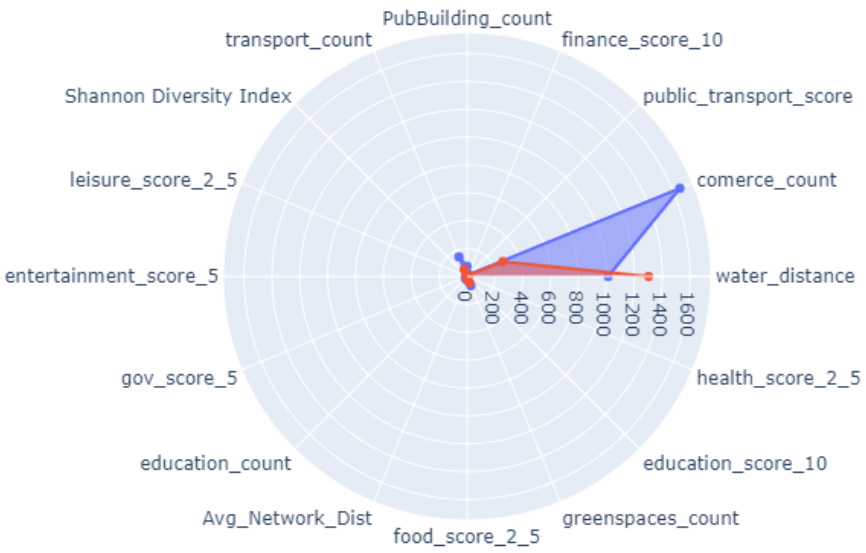
Name	Description	Calculation	Name	Description	Calculation
	closest green area			discarding 10 minutes	
Commerce Distance	Euclidean distance to the closest commercial establishment	Calculated the same way as the **Transport Distance**.	Public Transport Score	Sum of the distances, in minutes, to all public transportations in a 15-minute walk radius	Score = sum([1/(1+dist) for dist in a['distance']])
Public Building Distance	Euclidean distance to the closest public Building	Calculated the same way as the **Transport Distance**.	Public Transport Score 2.5	Sum of the distances to public transportations, discarding 2.5 minutes	Score_2.5 = sum([1/(1+dist+2.5) for dist in a['distance']])
Food Score	Sum of the distances, in minutes, to all establishments that sell food in a 15-minute walk	Score = sum([1/(1+dist) for dist in a['distance']])	Public Transport Score 5	Sum of the distances, in minutes to public transportations, discarding 5 minutes	Score_5 = sum([1/(1+dist+5) for dist in a['distance']])
Food Score 2.5	Sum of the distances to establishments that sell food, discarding 2.5 minutes	Score_2.5 = sum([1/(1+dist+2.5) for dist in a['distance']])	Public Transport Score 10	Sum of the distances, in minutes to public transportations, discarding 10 minutes	Score_10 = sum([1/(1+dist+10) for dist in a['distance']])
Food Score 5	Sum of the distances to establishments that sell food, discarding 5 minutes	Score_5 = sum([1/(1+dist+5) for dist in a['distance']])	Finance Score	Sum of the distances, in minutes, to all financial establishments in a 15-minute walk radius	**Score** calculated the same way as the previous amenities
Food Score 10	Sum of the distances to establishments that sells food, discarding 10 minutes	Score_10 = sum([1/(1+dist+10) for dist in a['distance']])	Finance Score 2.5	Sum of the distances to financial establishments, discarding 2.5 minutes	Calculated the same way as the previous amenities for S**core 2.5**
Leisure Score	Sum of the distances, in minutes, to all leisure establishments in a 15-minute walk radius	**Score** calculated the same way as the previous amenities	Finance Score 5	Sum of the distances to financial establishments, discarding 5 minutes	Calculated the same way as the previous amenities for S**core 5**
Leisure Score 2.5	Sum of the distances to leisure establishments, discarding 2.5 minutes	Calculated the same way as the previous amenities for S**core 2.5**	Finance Score 10	Sum of the distances to financial establishments, discarding 10 minutes	Calculated the same way as the previous amenities for S**core 10**

Name	Description	Calculation	Name	Description	Calculation
Leisure Score 5	Sum of the Distance to leisure establishments, discarding 5 minutes	Calculated the same way as the previous amenities for S**core 5**	Education Score	Sum of the distances, in minutes, to all educational establishments in a 15-minute walk radius	**Score** calculated the same way as the previous amenities
Leisure Score 10	Sum of the distance to leisure establishments, discarding 10 minutes	Calculated the same way as the previous amenities for S**core 10**	Education Score 2.5	Sum of the distances to educational establishments, discarding 2.5 minutes	Calculated the same way as the previous amenities for S**core 2.5**
Health Score	Sum of the distances, in minutes, to all health establishments in a 15-minute walk radius	**Score** calculated the same way as the previous amenities	Education Score 5	Sum of the distances to educational establishments, discarding 5 minutes	Calculated the same way as the previous amenities for S**core 5**
Health Score 2.5	Sum of the distances to health establishments, discarding 2.5 minutes	Calculated the same way as the previous amenities for S**core 2.5**	Education Score 10	Sum of the distances to educational establishments, discarding 10 minutes	Calculated the same way as the previous amenities for S**core 10**
Health Score 5	Sum of the distances to health establishments, discarding 5 minutes	Calculated the same way as the previous amenities for S**core 5**	Entertainment Score 5	Sum of the distances to entertainment establishments, discarding 5 minutes	Calculated the same way as the previous amenities for S**core 5**
Health Score 10	Sum of the distance to health establishments, discarding 10 minutes	Calculated the same way as the previous amenities for S**core 10**	Entertainment Score 10	Sum of the distance to entertainment establishments, discarding 10 minutes	Calculated the same way as the previous amenities for S**core 10**
Entertainment Score	Sum of the distances, in minutes, to all entertainment establishments in a 15-minute walk radius	**Score** calculated the same way as the previous amenities	Water Distance	Euclidean distance to the Tejo River	Water Distance = house_node.geometry.distance(river_node.geometry)
Entertainment Score 2.5	Sum of the distances to entertainment establishments, discarding 2.5 minutes	Calculated the same way as the previous amenities for S**core 2.5**			

Annex 1 – Table with all the calculated features, their names, description and formula.

Feature	Importance
Shannon Diversity Index	1423.94
Government Score	1033.45
Education Score 2.5	886.11
Government Score 2.5	847.97
Finance Score 2.5	526.85
Education Score 10	433.22
Health Score 2.5	418.10
Finance Score 10	331.98
Leisure Score 2.5	231.93
Food Score	207.02
Health Score 10	183.32
Leisure Score	131.63
Government Score 5	114.11
Food Score 2.5	111.49
Health Score	105.49
Leisure Score 10	59.53
Entertainment Score 5	52.59
Public Transport Score 2.5	39.63

Annex 2 – Lasso's selected features.



Annex 3- Radar chart of Avenidas Novas (blue) and Santa Clara (red) variables value.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa