ISBN: 978-981-18-5852-9
WCSE 2022 Spring Event: 2022 9th International Conference on Industrial Engineering and Applications
doi: 10.18178/wcse.2022.04.157

Impact of Commercial Risk on Real Estate Projects: An Application of Machine Learning to Real Estate

Felix Alberto Pacheco-Fuster[1], Juan Alejandro Ortega-Saco[2] and Franklin Cordova-Buiza[3,4+]

[1]Faculty of Economic Engineering, Statistics and Social Sciences, Universidad Nacional de Ingenieria, Lima, Peru

[2]Faculty of Industrial Engineering, Universidad Privada del Norte, Lima, Peru

[3]Reserch and Innovation Department, Universidad Privada del Norte, Lima, Peru

[4]Faculty of Business Sciences, Continental University, Huancayo, Peru

Abstract. The health crisis has marked a milestone in the digital transformation worldwide, influencing all processes in the various productive sectors of an economy. One example is the real estate industry, which had to digitize its legal, commercial, supervision, and analysis processes, etc., since it was affected by productive and social restrictions that had an impact on the increase in the rate of unemployment and interruption in the payment chain. As a result, it was assumed that the acquisition of a property in this context was a delayed need, understood as a commercial risk for real estate companies. However, there have been increases in several real estate markets around the world, showing that the impact of commercial risk in the real estate sector was on a smaller scale. In this sense, the use of big data and machine learning techniques have made it possible to identify and analyze such risks and the industrial engineer as a process professional is in charge of adapting these innovative tools in risk measurement. Due to these facts, this study has been motivated with the objective of measuring the impact of commercial risk in a developing economy, applying K-means technique, an unsupervised learning algorithm. The results showed that residential real estate projects did not demonstrate a significant impact of commercial risk, with an average increase of 8 months in the payback period, while massive projects did not have a statistically significant impact, increasing the payback period by an average of 13 months.

Keywords: Risk; Real Estate; Machine Learning; Clustering; Payback; Sales velocity.

1. Introduction

Restrictions on productive and social activities were common measures adopted by several countries to mitigate the spread of covid-19. However, there was a significant impact on the economies and their various productive sectors. One case was the real estate sector, which was exposed both at the supply level, with construction activities stopped, and at the demand level, with the increase in the unemployment rate, which represent risks at the business level. [1] defined business risk as the uncertainty that arises during the achievement of an objective, i.e., adverse circumstances, events or occurrences that impede the normal development of a company's activities that materialize in economic repercussions.

[2] explained that real estate markets are not ordinary markets; they are probably those that, to a lesser degree, comply with the conditions attributed to efficient and perfect competition markets. However, they recognized that the information asymmetry is a characteristic of the real estate market, since technical and legal specifications, among others, are better known to the seller than to the buyer, and for this reason, both buyers and sellers invest time and money in the search for information.

In contrast, [3] argued that large companies are better able to afford the costs of acquiring and processing relevant information, unlike small companies that may find it economically inviable. Therefore, he recognized that incomplete or asymmetric information addresses issues such as: adverse selection, moral hazard, decisions under uncertainty, credibility, among others. It is also possible to associate commercial risk, since it is materialized in the probability of not making sales as expected.

+ Corresponding author. Tel.: +51-947186112
E-mail address: franklin.cordova@upn.edu.pe

In September 2020 [4] published an article on an unusual real estate boom in the midst of the worst economic crisis of the last decades, reflected in the increase of the price per square meter in several countries. This article explained that this unusual real estate boom occurs when organizations such as the International Monetary Fund (IMF) speak of a global economic slowdown.

Subsequently, in May 2021 [5] published an article on the continuation of the global house price boom during 1Q21 in Europe, the U.S., Canada and parts of Asia-Pacific, where he explained that low interest rate and monetary easing were the main cause. In addition, he reported that the real price (inflation-adjusted price) of housing increased in 43 of the 57 real estate markets in the world with housing statistics on record.

While in various parts of the world there are price increases and real estate booms, considering a period of global health crisis, these facts motivate to question whether the same impact occurs in any market in Latin America (LATAM) and if such information is symmetrical among the agents of that market. That is why the question arises: What is the impact of commercial risk on the payback period of real estate projects in the regions of Lima and Callao, Peru?

2. Theoretical framework

[6] indicated that real estate analysts use various proxies/indicators to analyze changes in demand, most of which are absorption measures. In addition, they stated that, from the point of view of the real estate developer, due to the durability of real estate, it is appropriate to measure marginal changes in demand rather than aggregate demand. Thus, whether used properly or misused, indicators of marginal changes in demand include gross, net and average absorption.

[7] defined Mean Time on Market as the average number of days on the market of real estate properties actually sold within an area, measured monthly, quarterly or annually. This mean is also known in the real estate industry as 'Sales Velocity' (α) which is obtained by dividing by the stock sold (Svt,T) over a period of time (k), where (k) is the time interval measured between a current time or cut-off date (T) and a previous time or last cut-off date (t), equation 1.

$$\alpha = Sv_{T,k} / (T - t) \quad (1)$$

[8] published an article on how to calculate the absorption rate where he defined supply as the available properties (St), demand as the properties sold (Svt,T) and the relationship between them as the absorption rate. Thus, he defined a second expression for the absorption rate that is equivalent to the absorption term (β) - or Inverse Absorption - by dividing the total number of available properties (ST) by the sales velocity, equation 2.

$$\beta=(S_T) / ((Sv_{t,T})/(T-t)) \quad (2)$$

[9] in their study on how CFOs make capital budgeting and capital structure decisions, they found that apart from NPV and IRR, the payback period was the most commonly used capital budgeting technique. This result, however, was surprising to them as there are financial theories that highlight shortcomings of the payback criterion since it ignores the time value of money and the value of cash flows beyond the cut-off date and this is often arbitrary.

[10] defined the traditional payback period as the expected time to recover the original investment, i.e., it is an expression of the year immediately preceding the full recovery of the initial investment (At) plus the result of dividing the amount of the initial investment that is not recovered at the beginning of the payback year RIt) and the Total cash flow generated during the payback year, equation 3.

$$PR=(A_{t-1}) + ((RI_t)/(FC_t)) \quad (3)$$

The payback period for a real estate project under development is the sum of the period since sales began (Fi) to the cut-off date (Fa) and the absorption period at the cut-off date, equation 4.

$$PRP=(Fa-Fi+\beta.30) / 30 \quad (4)$$

3. Methodology

3.1. Research design

According to its approach, it is quantitative, since it uses quantitative methods and techniques and therefore is related to measurement [11]. According to its level or knowledge pursued, it is basic, pure or fundamental, since it provides the background for applied or technological research [11]. According to its scope, it is descriptive, since the interest is to characterize phenomena, situations or events, indicating their most distinctive or differentiating features [12]. According to the strategy or design, it is non-experimental or observational, since the intention is not to demonstrate cause-effect relationships between variables, but to observe the phenomena as they occur in their natural context [12]. According to the planning of measurements, it is retrospective, since the data have been collected previously with existing records [13]. According to the number of measurements over time, it is cross-sectional or sectional, since information is obtained from each object of study only once at a certain time [14].

3.2. Sample

- Population: ~95% of real estate projects in the Department of Lima, including the Constitutional Province of Callao. This population considers real estate projects by type of property (single-family or houses, multi-family or apartments, and land or lots) and by type of financing (self-financed, mainly by shareholders and/or sales, and financed, mainly by entities of the Peruvian financial system).
- Sample: deterministic sampling by exclusion, being the real estate projects in the regions of Lima and Callao, of multifamily type (excluding single-family houses or lots) and that are financed by any entity of the Peruvian financial system (excluding self-financed).

3.3. Data Collection

The database was obtained from the Incoin Analytcs platform owned by Tinsa, a multinational provider of real estate valuation and advisory services with more than 30 years of experience in the real estate sector, headquartered in Spain. Its main clients are financial institutions in the countries where it operates, companies from the sector and public institutions.

The information of Tinsa has a scope of ~95% of the total real estate market data and is present in more than 25 countries. At international level, it is present in Europe (Spain, Portugal, Holland, Morocco and Belgium) and in LATAM (Mexico, Colombia, Ecuador, Peru, Argentina and Chile). In Peru, it has information on the real estate market in Lima, Callao, La Libertad, Lambayeque, Piura, Arequipa and Ica.

4. Problems and hypothesis

4.1. Problems

The first premise of this study is that the slowdown of the economy during the 2020, reducing the demand for credit, would have influenced: i) the reduction in interest rates, motivated by the competition between financial institutions to place liquidity; and ii) the reduction in prices of apartments, motivated by the competition between real estate companies.

Thus, the second premise focuses on how the demand in the real estate market was affected at the different socioeconomic levels, i.e.: i) socioeconomic levels with higher incomes would have shown greater prudence to invest or purchase, affected by changes in the pace of work and personal life; and ii) socioeconomic levels with lower incomes would have been affected by the unemployment increase.

However, according to data from the Superintendency of Banking, Insurance, and Private Pension Fund Administrators of Peru (SBS), during 2020 the Peruvian market for new mortgage loans reached an annualized growth of +5% in soles and +140% in dollars. In addition, according to the Peruvian Chamber of Construction (Capeco), real estate companies projected a 9% growth in their sales levels for 2021.

Considering the growth in the Peruvian mortgage market and the assumptions of credit contraction and impact on demand in the real estate market, it is suggested that commercial risk would have worsened by

having a growing, but increasingly limited real estate market, causing significant variations in the estimated payback periods for the development of the projects. Then, how to measure the impact of commercial risk in real estate projects in Lima and Callao?

4.2. Hypothesis

Usually, the viability analysis of a financed project includes a sales velocity that projects a cash flow that allows covering the total investment and recovering the contribution and profit of the project, in such a way that the investment is expected to be recovered in an estimated period of time, that is, the payback period. The probability of not selling as expected would extend this payback period, affecting the profitability initially expected by the stakeholders. Thus, a higher variability in the central parameters of the payback period will reflect the impact of commercial risk. This study aims to find the variation in the median of the payback period for real estate projects at the end of Q4 over the last three years, based on the following hypothesis:

H0: The medians of the payback period do not show significant differences (m1=m2)

H1: The medians of the payback period do show significant differences (m1 != m2)

5. Results

5.1. Sizing Bias - Unsupervised Learning Algorithm (K-means)

The real estate market has projects with different magnitudes or density, for example, there are multifamily projects with 3 to more than 120 real estate units per building, so it would not make sense to compare these projects in a single measurement since there would be a sizing bias, which should be addressed to avoid atypical values common to each group of projects that share similar dimensions.

In order to deal with this sizing bias, the k-means algorithm is used as a grouping technique. This is an unsupervised machine learning technique. This algorithm clusters data by trying to separate samples in *'n'* groups of equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares.

The aim is to identify centroids of several groups in the sample dispersion, in such a way that those groups where the distance between their centroids is the greatest possible are identified, measured with the parameter 'betweenss', which is equivalent to the sum of squares of the distance of the centroids between groups, and the distance of each record with regard to the centroid of its group is the smallest possible, measured with the parameter 'withinss' which is the sum of squares of the distances between each record and the centroid of the group to which it belongs.

A total of eight algorithms were trained considering 21, 15, 16, 13, 9, 6, 5, 4 and 3 numerical variables such as stocks, areas, prices, among others. Table I shows the results obtained from the eight algorithms. Although the algorithm with more variables (21) has a good performance in terms of distance between groups, 'betweenss', the same is not true for the distances within each group. On the contrary, the algorithm with fewer variables (3) has a good performance within each group, 'withinss', but not with the distance between groups.

Table 1: Results of k-means algorithm

N° Variables	Betwenss	Withinss	Withinss (in thousands)				
			G1	G2	G3	G4	G5
21	187 952	292 465	64.8	74.9	57.7	55.8	39.2
16	164 773	201 258	51.0	43.4	41.5	38.7	26.7
13	145 611	151 790	15.3	33.7	35.5	31.8	35.6
9	120 362	85 530	19.9	12.6	17.4	19.3	16.4
6	84 567	52 695	6.6	9.2	15.6	8.3	13.0
5	73 691	40 694	20.7	9.3	3.2	3.4	4.1
4	59 508	31 999	14.5	7.6	2.6	3.2	4.1
3	47 911	20 720	6.3	5.0	2.4	2.8	4.1

Source: Tinsa – Incoin Analytics. Own elaboration.

5.2. Validation of the algoithm

There are three approaches to investigate the validity of clusters [16]. The first is based on external criteria. This implies that the results of a clustering algorithm based on a pre-established structure are evaluated, which is imposed on a dataset and reflects our intuition about the clustering structure of the dataset [17]. The second approach is based in internal criteria, where are evaluated in terms of quantities involving the vectors of the data set [17].

For choosing the algorithm, the model with the smallest increase in the sum of squares between groups (vertical line, ‘elbow’ technique) is taken as a reference. From the results obtained in Figure 1, it can be observed that the algorithms with 9 and 6 variables would present a better performance if more groups are considered. Although the 5-variable algorithm achieves an adequate performance with 5 groups, it is observed that one of the groups has an average of 42 units of initial stock, but has a deviation of 44 units, which would be inconsistent since we would obtain negative values.

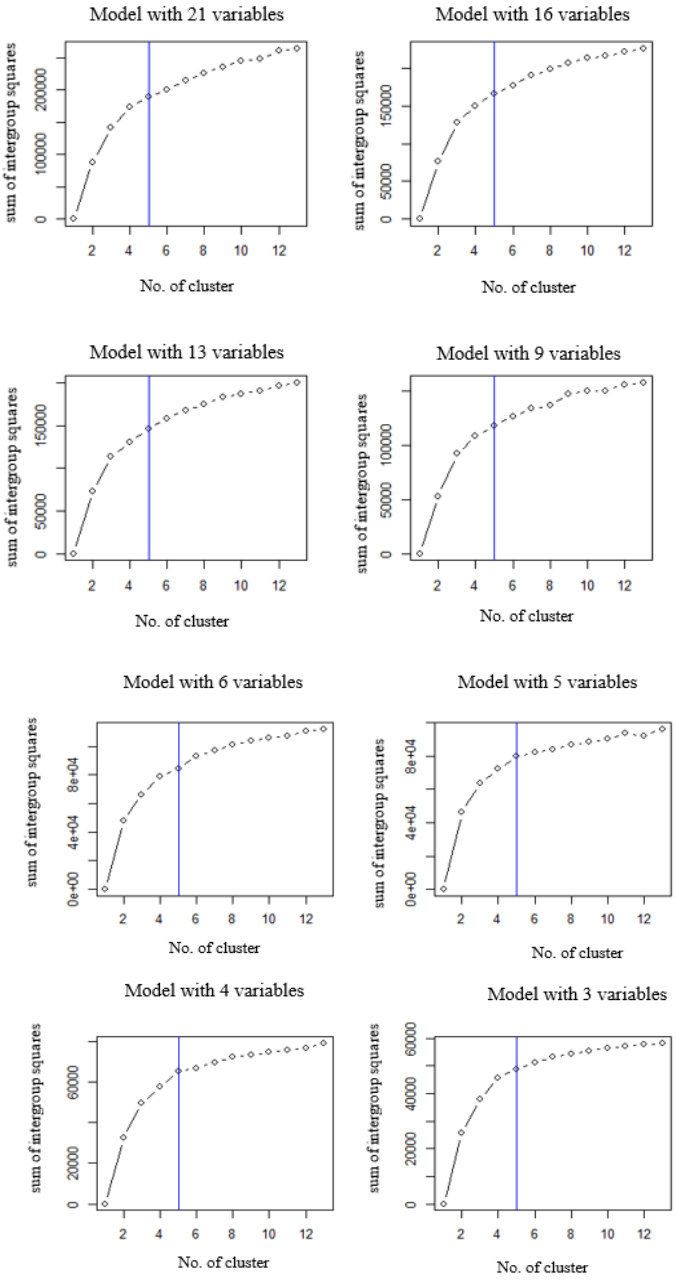

Fig. 1: Algorithms

As the treatment is the sizing bias, the most consistent algorithm concerning to the standard deviation is chosen, where the attribute 'Initial Stock' will be the classifier variable between the groups. The algorithm with 6 variables has higher parameter 'betweenss' and lower 'withinss'. Finally, the groups ranges are

determined according to external criteria, in other words, the author's economic approach and/or the average and standard deviation of the 'Initial Stock' for each group, table 2.

Table 2: Determination of ranges by cluster

N° groups	X	σ	X - σ	X+ σ	Classif.	Range
G1	25	21	4	46	Micro	3 – 25
G2	209	83	126	292	Big	176 – 275
G3	758	487	271	1245	Massive	276 – more
G4	78	31	47	109	Small	26 – 75
G5	174	74	100	248	Midsized	76 – 175

Source: Tinsa – Incoin Analytics. Own elaboration

The elections of the established groups order the variables price per apartment, roofed area and average sales velocity appropriately and with economic sense. Table 3 shows how Micro projects are mainly aim at higher income socioeconomic levels, reaching the highest market areas and prices. The Small and Midsized projects, mainly aimed at middle-income socioeconomics levels, are located in districts in urban growth. Big and Massive projects, mainly aimed at social housing, access the economic incentives provided by the government.

Table 3: Sort validation

N° grupos	Average Price	Roofed area	Monthly velocity
Micro	PER 861 625	120 m^2	0.44 unit/month
Small	PER 584 809	89 m^2	1.12 unit/month
Midsized	PER 400 396	73 m^2	2.78 unit/month
Big	PER 286 810	64 m^2	4.06 unit/month
Massive	PER 251 299	67 m^2	8.55 unit/month

Source: Tinsa – Incoin Analytics. Own elaboration

5.3. Absorption period by segment

Finding the payback period requires estimating the period in which the available stock would be sold, the absorption period, so that considering a minimum velocity equal to zero would not make sense. For this reason, a minimum velocity is determined for each classification equivalent to the maximum between 0.1000 and the percentile 15 in each group (j), equation 5 and table 4.

$$\min(\alpha_j) = \max\{P_{15}; 0.00\} \quad (5)$$

Table 4: Minimum velocity

j	micro	small	midsized	big	massive
α_j	0.1000	0.1000	0.3000	0.5882	0.5503

Source: Tinsa – Incoin Analytics. Own elaboration

With the minimum velocity [mín(α_j)], the absorption period for each project is estimated (i) according to its group (j) by dividing the stock available for sale (Sd_{ij}) and the sales velocity of each project according to its segment, conditioned to the minimum sales velocity if it were zero, equation 6.

$$\beta_{ij} = Sd_{ij} / \text{máx}(\alpha_{ij}; \text{mín}(\alpha_j)) \quad (6)$$

With the absorption period, the payback period is estimated, i.e., the total time in which the total number of units of each project will be sold. This payback period is expressed as the sum of the months elapsed from the date when sales of a project began until the update date of each record and the absorption period of each project, equation 7.

$$RP_{ij} = (Fa_{ij} + \beta_{ij}.30 - Fi_{ij}) / 30 \quad (7)$$

5.4. Medians of the payback period

Table V shows the calculation of medians of the payback period by groups, measured in the fourth quarters from 2017 to 2020.

Table 5: Medians of the payback period by cluster

Quarter	Cluster				
	micro	small	midsized	big	massive
4Q 2017	39.63149	47.84354	46.60367	59.17434	77.33333
4Q 2018	45.86700	44.43333	43.27113	57.94993	94.74998
4Q 2019	43.60000	40.20094	44.93333	53.92410	85.75519
4Q 2020	49.45956	49.60000	54.10050	73.12805	91.60006

Source: Tinsa – Incoin Analytics. Own elaboration

5.5. Mann-Whitney U test

Commercial risk is considered as the variation of the medians of the payback period of the multifamily projects financed in Lima, i.e., if 50% of the population shows statistically significant variation in the payback period, it is interpreted that the existence of commercial risk impacts on the payback period.

Table VI shows the p-values and decision results of the payback period for each group. The tests were done on pairs of annualized quarters, where the contrast decision for the null hypothesis will be either Not Rejected (NSR) or Rejected (SR).

Table 6: Test U de Man Whitney

Quarter	Cluster				
	micro	small	midsized	big	massive
4Q17-4Q18	0.1383 NSR	0.9801 NSR	0.2755 NSR	0.7780 NSR	0.8776 NSR
4Q18-4Q19	0.8657 NSR	0.0650 NSR	0.6334 NSR	0.9891 NSR	0.8066 NSR
4Q19-4Q20	0.0109 SR	0.0000 SR	0.0000 SR	0.0349 SR	0.1474 NSR
4Q18-4Q20	0.0057 SR	0.0101 SR	0.0000 SR	0.0319 SR	0.4049 NSR

Source: Tinsa – Incoin Analytics. Own elaboration

6. Discussion

Regarding the sizing bias, the results indicate that projects classified as micro and small would have a sales velocity slower than 0.1000 units per month. In this sense, 0.1000 units per month is considered as the minimum velocity. This is due to the fact that these classifications are usually composed of residential projects addressed to higher socioeconomic levels, which offer the highest prices in the market. Although [5] explained a worldwide boom in housing prices, in Peru, the results indicate that the projects with higher prices would show a slower sales velocity, even slower than selling one unit every ten months, on average.

Regarding the medians of the payback period, it is observed that, from 4Q2017 to 4Q2018, the payback period decreased by an average of 2.7 months in the small, midsized and large groups. However, in the micro and massive groups, the payback period increased by an average of 11.8 months. In contrast, from 4Q2018 to 4Q2019, it decreased by an average of 4.8 months in almost all groups, except for the midsized group, which increased by 1.6 months.

[4] reported an unusual real estate boom in the midst of the worst economic crisis of the last decades. However, the results indicate that from 4Q2019 to 4Q2020 the payback period increased by an average of 9.9 months in all groups. It is possible to imagine that such variations respond to the measures imposed against the advance of covid, affecting the various productive sectors and contracting domestic demand.

7. Conclusions

From 4Q17 to 4Q18, the null hypothesis is not rejected in any of the groups. Thus, no commercial risk that impacts the payback period of the real estate projects has been identified. Likewise, considering that the medians of the payback period for the ‘micro’ and ‘massive’ groups increased by +6 months and +11 months, respectively, these increases do not present sufficient statistical evidence to reject the null hypothesis.

Similarly, from 4Q18 to 4Q19 the results indicate that the null hypothesis is not rejected in any of the groups, identifying that there would not be a commercial risk that would significantly impact the payback

period. Contrasting the two previous tests, the results from 4Q19 to 4Q20 indicate that the null hypothesis is rejected in almost all groups. In this sense, a commercial risk that significantly impacts the payback period is identified. However, in the case of the 'massive' group, the results indicate that the null hypothesis is not rejected, identifying that the commercial risk does not have a significant impact on the payback period.

It is important to note that from 4Q18 to 4Q19 the medians decreased in almost all groups, inferring that 2019 would have been a better performing year compared to 2018. However, making a comparison to a year stressed by the health crisis could skew the results. To verify this assumption, the hypothesis of equality of medians between 4Q18 and 4Q20 was contrasted.

The results obtained were similar: the null hypothesis was rejected in almost all groups. In this sense, it is identified that commercial risk had a significant impact on the payback period. Similarly, it is confirmed that, for the 'massive' group, the null hypothesis is not rejected, so that commercial risk did not have a significant impact on the payback period.

The results obtained were similar: the null hypothesis was rejected in almost all groups. In this sense, it is identified that commercial risk had a significant impact on the payback period. Similarly, it is confirmed that, for the 'massive' group, the null hypothesis is not rejected, so that commercial risk did not have a significant impact on the payback period.

8. References

[1] ISOTools Excellence. (2019). "El valor de la gestión de riesgos en las organizaciones" [The value of risk management in organizations], *ISOTools* [Internet], pp. 4, Available: www.isotools.org.

[2] G. Aranda, M. Castillo, & A. Rodriguez. (2003). "El mercado de vivienda y su enfoque neoinstitucional" [The housing market and its new institutional approach], *Análisis Económico* [Internet], vol. XVIII, no. 39, pp. 287-301, Available: https://www.redalyc.org/articulo.oa?id=41303913.

[3] J. Ayala. *Instituciones y Economía una introducción al Neoinstitucionalismo* [Institutions and Economics: an introduction to New institutionalism.], Mexico D.F.: Fondo Cultura Económica, 1999, pp. 396.

[4] BBC. (Sep. 2020). "Cómo se explica el insólito 'boom' inmobiliario en medio de la peor crisis económica de las últimas décadas" [How to explain the unusual real estate boom in the midst of the worst economic crisis of the last decades], BBC News Mundo [Intenet], Available: https://www.bbc.com/mundo/noticias-54035630.

[5] M. Montagu-Pllock. (May 2021). "Q1 2021: Global house price boom continues amazingly strong, led by Europe, U.S., Canada and parts of Asia-Pacific", *The Global Property Guide* [Intenet], Available: https://www.globalpropertyguide.com/investment-analysis/Q1-2021-Global-house-price-boom-continues-amazingly-strong-led-by-Europe-US-Canada-and-parts-of-Asia-Pacific.

[6] M. Natsvaladze & M. Beraia. *Real Estate Economics*. Tbilisi, 2014

[7] N. Miller & M. Sklarz. (1986). "A note on leading indicators of housing market price trends", *The Journal of Real Estate Research*, pp. 99-109.

[8] M. Anaya. (Aug. 2020). "Como calcular la tasa de absorción" [How to calculate the absorption rate], *Land & Building magazine* [Internet], Available: https://landandbuilding.com/blog/2020/08/10/como-calcular-la-tasa-de-absorcion/.

[9] J. Graham & H. Campbell. (2002). "How do CFOs make capital budgeting and capital structure decisions?". *Journal of applied Corporate Finance* [Online], pp. 8-23.

[10] S. Besley & E. Brigham. FINC: Finanzas Corporativas [FINC: Corporate Finance], 4th ed. Mexico D.F.: Cengage Learning Editores SA, 2016.

[11] H. Ñaupas, M. Valdivia, J. Palacios & H. Romero. *Metodología de la Investigación Cuantitativa-Cualitativa y Redacción de la Tesis* [Quantitative-Qualitative Research Methodology and Thesis Writing], 5th ed. Bogota: Ediciones de la U, 2018.

[12] G. Mousalli-Kayat. *Métodos y Diseños de Investigación Cuantitativa* [Quantitative Research Methods and Designs]. *Merida*, 2015.

[13] R. Jimenez. *Metodología de la Investigación: Elementos Básicos para la Investigación Clínica* [Research Methodology: Basic Elements for Clinical Research], La Habana: Ciencias Médicas, 1998.

[14] C. Bernal. *Metodología de la Investigación* [Research Methodology]. Bogota: Pearson Educación de Colombia Ltda, 2010.

[15] Pedragosa et al. *Scikit-learn: Machine Learning in Python*, JMLR 12, 2011, pp. 2825-2830.

[16] S. Theodoridis, K. Koutroubas (1999). *Pattern recognition*, Academic Press.

[17] M. Halkidi, M. Vazirgiannis, I. Batistakis (2000). Quality scheme assessment in the clustering process. In: Proc. PKDD (Principles and Practice of Knowledge in databases), Lyon, France, Lecture Notes in Artificial Intelligence, Spring-Verlag, Vol. 1910, pp. 265-279