

19-21 de Octubre 2022 | Granada

INTERNATIONAL CONFERENCE ON REGIONAL SCIENCE

Challenges, policies and governance of the territories in the post-covid era

Desafíos, políticas y gobernanza de los territorios en la era post-covid

XLVII REUNIÓN DE ESTUDIOS REGIONALES
XIV CONGRESO AACR



RESUMEN AMPLIADO

Título: Precio de la vivienda en Barcelona antes y después de la COVID, una aproximación mediante *Machine Learning*

Autores y e-mail de todos ellos: Mora-Garcia, Raul-Tomas; Cespedes-Lopez, Maria-Francisca; Perez-Sanchez, Vicente-Raul
rtmg@ua.es, paqui.cespedes@ua.es, raul.perez@ua.es

Departamento: Edificación y Urbanismo

Universidad: Universidad de Alicante

Área Temática: 13. Ordenación del territorio, urbanismo y vivienda

Palabras Clave: Precio de la vivienda, Valoración masiva, Machine Learning, Hiperparámetros, Evaluación del modelo, Barcelona.

Clasificación JEL: R31 - Oferta y mercados de la vivienda

Resumen

1. Introducción

Los efectos de la pandemia, producto del COVID-19, han incidido en todos los ámbitos, el sanitario, el económico, el laboral, el productivo y el social, entre otros. El mercado de la vivienda no ha sido ajeno a estos efectos devastadores, pero los datos de las series históricas sugieren que el impacto será mucho menor que el ocasionado en la crisis económica iniciada en el último trimestre de 2007 y que duró hasta 2014.

El mercado de la vivienda en España se ha visto afectado por la crisis sanitaria fruto de la pandemia. Desde 2014 se estaba produciendo un crecimiento expansivo del sector inmobiliario, estabilizándose durante el 2019. Existen grandes dudas sobre cómo va a evolucionar el mercado de la vivienda a corto, medio y largo plazo.

Esta investigación tiene por objeto estudiar la evolución de los precios de las viviendas en la ciudad de Barcelona, antes y durante la pandemia. Se pretende cuantificar el impacto en el precio de venta a lo largo del tiempo, mediante un estudio con datos de corte transversal agrupados (*pooled cross section*).

Existen estudios que han investigado la evolución de los precios de la vivienda en otras situaciones de pandemia como las ocurridas por la Peste en Ámsterdam (siglo XVII) y el cólera en París (siglo XIX) (Francke & Korevaar, 2021), y la epidemia del Síndrome Respiratorio Agudo Severo (SARS) en Hong Kong en el 2003 (Wong, 2008). Otros estudios han abarcado los años anteriores a la crisis financiera de 2008, donde se produjo un crecimiento expansivo en el mercado inmobiliario (Nicodemo & Raya, 2012; Taltavull & McGreal, 2009). Otros estudios se han centrado en los efectos de la crisis financiera de 2008 en la economía y el mercado inmobiliario (Duca et al., 2010; Neal & García-Iglesias, 2013; Royo, 2009). Desde el Banco de España se han realizado publicaciones con información del mercado de la vivienda español anterior a la crisis sanitaria (Banco de España, 2020) y posterior a ella (Alves Álvarez & San Juan del Peso, 2021). Existen otros trabajos donde se investigan los efectos de la pandemia en diferentes regiones, como EE.UU (Li & Zhang, 2021; Ouazad, 2021), la zona Euro (Battistini et al., 2021), Wuhan (China) (Cheung et al., 2021), Australia (Hu et al., 2021) o Turquía (Kartal et al., 2021).

El análisis temporal de los casos COVID-19 declarados diariamente a la Red RENAVE (ISCIH, 2021) muestran varias olas en España, coincidiendo con los periodos de vacaciones estivales (julio de 2020 y 2021), así como el gran impacto de los contagios entre diciembre de 2020 y febrero de 2021, resultado de las festividades navideñas.

Por otro lado, los algoritmos de aprendizaje automático (*Machine Learning*, en adelante ML) se están implementando con más frecuencia para la valoración masiva inmobiliaria (Antipov & Pokryshevskaya, 2012; Banerjee & Dutta, 2017; Čeh et al., 2018; De Aquino Afonso et al., 2019; Hu et al., 2019; Kok et al., 2017; Park & Bae, 2015).

En la valoración masiva se utilizan procedimientos estandarizados donde se recopilan datos de ofertas inmobiliarias para realizar estimaciones de valor en grandes conjuntos de inmuebles, pudiéndose garantizar que las valoraciones se realicen de forma homogénea e imparcial.

En esta investigación se utilizan varios algoritmos de aprendizaje automático para desarrollar un modelo de predicción del precio de la vivienda en un caso de estudio de la

ciudad de Barcelona. Se describe la metodología utilizada en todo el proceso, desde la adquisición de datos hasta el despliegue del modelo.

2. Materiales y método

Este estudio se ha centrado en la ciudad de Barcelona, capital de provincia que alberga el 28,6% del total de población censada en la provincia del mismo nombre (INE, 2021). La ciudad de Barcelona muestra una gran diferencia espacial en la distribución de los precios de la vivienda, concentrándose los precios más altos en la zona norte del distrito de “Les Corts”, zona sureste del distrito “Sarrià-Sant Gervasi” y barrio de “la Dreta de l'Eixample” en el distrito de “l'Eixample”.

La información de precios de oferta inmobiliaria se ha extraído de un portal web inmobiliario, recolectando precios de oferta y características tanto de las viviendas como de los edificios. Se recogieron datos sobre el precio de venta ofertado, las características de la vivienda (tipología constructiva, superficie construida, número de dormitorios, baños y aseos, instalaciones, etc.), características del edificio (plaza de garaje, ascensor, piscina, trastero, obra nueva), características de ubicación (coordenadas geográficas) y temporales (trimestre). Para modelar el efecto del tiempo, se han creado variables ficticias que identifican si un determinado inmueble se comercializa en ese momento temporal.

Durante los años 2019, 2020 y 2021 se han recogido de un portal inmobiliario cerca de 118.000 ofertas de venta de viviendas plurifamiliares. Se realizó un proceso donde se buscaban nuevos precios ofertados no identificados en las descargas anteriores, se verificaba si los inmuebles existentes seguían en venta o si habían variado el precio, y se daban de baja los inmuebles no encontrados. De esta manera ha sido posible documentar, para cada inmueble, el periodo que ha estado ofertado en el mercado, así como sus variaciones de precio a lo largo del tiempo.

La información aportada por los anunciantes a veces está incompleta o defectuosa, por lo que existen datos con inconsistencias que deben revisarse. Se identificaron valores poco probables en las variables cuantitativas, como en el tamaño de la vivienda, el número de dormitorios o en los precios. Para su identificación se realizó un análisis de valores atípicos univariados, descartando inmuebles con valores mayores o menores de cinco desviaciones estándar en todas las variables cuantitativas. Una vez completada la base de

datos, se excluyeron aquellos casos en los que faltaban datos para algunas características relevantes (como superficie, nº de dormitorios y baños, y coordenadas geográficas).

3. Metodología

En esta investigación se utilizan varios algoritmos de aprendizaje automático para desarrollar un modelo de predicción del precio de la vivienda en un caso de estudio de la ciudad de Barcelona. Para ello se han abordado todos los pasos previstos en el flujo de trabajo habitual de *Machine Learning*: Preparación de los datos, Ingeniería de características, Selección del modelo, Optimización de hiperparámetros, Evaluación del modelo, Interpretación del modelo y Despliegue del modelo.

- 1) Preparación de los datos (*Data preparation*). Proceso para la limpieza y transformación de los datos sin procesar previos a realizar tareas de minería de datos. En este proceso se realizan tareas para dar formato a los datos (agrupación, discretización, binarización), enriquecerlos con información de otras fuentes, identificación y tratamiento de los datos atípicos y faltantes, transformación de las variables (estandarización, normalización).
- 2) Ingeniería de características (*Feature engineering*). En este proceso se utilizan técnicas para extraer características de datos sin procesar mediante técnicas de minería de datos (*feature creation, feature extraction y feature selection*). Estas técnicas permiten mejorar el rendimiento de los algoritmos de aprendizaje automático. A partir de varios modelos de referencia (línea base) se valora la posibilidad de codificar las variables categóricas; generar nuevas características como agregación, transformación o combinación de otras existentes; creando nuevas características que resuman las variables originales (análisis de componentes principal y discriminante, *auto-encoding*); y seleccionar las características más relevantes mediante selección recursiva (*sequential feature selection, recursive feature elimination*).
- 3) Selección del modelo (*Model selection*). En esta tarea se entrenan varios modelos candidatos para estimar el poder predictivo de los mismos y determinar qué algoritmos tienen mejor desempeño. Para la selección del modelo se utilizan métricas de error (MAE, RMSE, MAPE, etc), bondad de ajuste (R^2) u otros índices de selección que tienen en cuenta la bondad de ajuste y la complejidad del modelo (BIC y AIC).

- 4) Optimización de hiperparámetros (*Hyperparameter optimization*). Con esta fase se pretende mejorar la bondad de ajuste del modelo y/o minimizar los errores de predicción. Una vez identificado el mejor o mejores algoritmos, se procede a evaluar la incidencia de los hiperparámetros con la finalidad de ajustarlos mediante técnicas de validación cruzada (*Cross Validation*). Es habitual utilizar un subconjunto de datos de entrenamiento para aplicar una variedad de combinaciones de hiperparámetros y evaluar su desempeño en otro subconjunto de datos de validación.
- 5) Evaluación del modelo (*Model evaluation*). En esta fase se evalúa el modelo, identificando el que mejor desempeño ha tenido para predecir la variable dependiente. Además, se estudia la posible existencia de sobreajuste (*Overfitting*) obtenidos en la fase anterior mediante la técnica de validación cruzada y subconjuntos de entrenamiento y prueba.
- 6) Interpretación del modelo. Se utilizan herramientas que permiten identificar las características más importantes del modelo mediante enfoques globales y locales, así como técnicas gráficas de visualización de las predicciones. Para ello, se pueden utilizar gráficos de dependencia parcial (*PDP Partial Dependence Plot*), gráficos de expectativa condicional individual (*ICE Individual Conditional Expectation*), gráficos de efectos locales acumulados (*ALE Accumulated Local Effects*) y valores de Shapley (*SHAP SHapley Additive exPlanations*), entre otros.
- 7) Despliegue del modelo (*Deployment*). En esta fase se desarrolla una plataforma web para poner el modelo en producción, permitiendo realizar predicciones a partir de unos datos iniciales.

4. Conclusiones

Se ha comprobado la viabilidad de utilizar algoritmos de aprendizaje automático para estimar el precio de la vivienda en la ciudad de Barcelona, ya que mejoran sustancialmente los resultados de métodos tradicionales como la regresión lineal. Además, estos algoritmos se adaptan mejor a las no linealidades de las variables inmobiliarias. Ha sido posible confeccionar un código fuente en Python que abarca todo el proceso de ciencia de datos: Preparación de los datos, Ingeniería de características, Selección del modelo, Optimización de hiperparámetros, Evaluación del modelo, Interpretación del modelo y Despliegue del modelo.

Referencias

- Alves Álvarez, P. A., & San Juan del Peso, L. (2021). El impacto de la crisis sanitaria del COVID-19 sobre el mercado de la vivienda en España. *Boletín económico del Banco de España*, 2021(2), 1-15.
<https://repositorio.bde.es/handle/123456789/16540>
- Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, 39(2), 1772-1778. <https://doi.org/https://doi.org/10.1016/j.eswa.2011.08.077>
- Banco de España, Dirección General de Economía y Estadística. (2020). *El mercado de la vivienda en España entre 2014 y 2019* (Vol. 2013). Banco de España.
<https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSerias/DocumentosOcasionales/20/Fich/do2013.pdf>
- Banerjee, D., & Dutta, S. (2017, 21-22 Sept. 2017). Predicting the housing price direction using machine learning techniques. 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, India.
- Battistini, N., Falagiarda, M., Gareis, J., Hackmann, A., & Roma, M. (2021). The euro area housing market during the COVID-19 pandemic. *European Central Bank: Economic Bulletin*, 2021(7), 115-132.
<https://www.ecb.europa.eu/pub/pdf/ecbu/eb202107.en.pdf>
- Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. 7(5), 168. <https://doi.org/https://doi.org/10.3390/ijgi7050168>
- Cheung, K. S., Yiu, C. Y., & Xiong, C. (2021). Housing Market in the Time of Pandemic: A Price Gradient Analysis from the COVID-19 Epicentre in China. *Journal of Risk and Financial Management*, 14(3), 108.
<https://doi.org/https://doi.org/10.3390/jrfm14030108>
- De Aquino Afonso, B. K., Carvalho Melo, L., Dihanster, W., Sousa, S., & Berton, L. (2019). *Housing Prices Prediction with a Deep Learning and Random Forest Ensemble* (I Engineering Education for the Future (EEF 2019). Data Science Challenge, Issue. <https://sol.sbc.org.br/index.php/eniac/article/view/9300/9202>
- Duca, J. V., Muellbauer, J., & Murphy, A. (2010). Housing markets and the financial crisis of 2007–2009: Lessons for the future. *Journal of Financial Stability*, 6(4), 203-217. <https://doi.org/https://doi.org/10.1016/j.jfs.2010.05.002>
- Francke, M., & Korevaar, M. (2021). Housing markets in a pandemic: Evidence from historical outbreaks. *Journal of Urban Economics*, 123, 103333.
<https://doi.org/https://doi.org/10.1016/j.jue.2021.103333>
- Hu, L., He, S., Han, Z., Xiao, H., Su, S., Weng, M., & Cai, Z. (2019). Monitoring housing rental prices based on social media: An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies. *Land Use Policy*, 82, 657-673.
<https://doi.org/https://doi.org/10.1016/j.landusepol.2018.12.030>
- Hu, M. R., Lee, A. D., & Zou, D. (2021). COVID-19 and Housing Prices: Australian Evidence with Daily Hedonic Returns. *Finance Research Letters*, 43, 101960.
<https://doi.org/https://doi.org/10.1016/j.frl.2021.101960>
- INE, Instituto Nacional de Estadística. (2021). *Padrón Población por municipios. Cifras oficiales de población de los municipios españoles: Revisión del Padrón Municipal*.

- https://www.ine.es/dyngs/INEbase/categoria.htm?c=Estadistica_P&cid=1254734710990
- ISCIII, Instituto de Salud Carlos III. (2021). *COVID-19 - Documentación y datos* (cneccovid.isciii.es). Retrieved 09/2021 from <https://cneccovid.isciii.es/covid19/#documentaci%C3%B3n-y-datos>
- Kartal, M. T., Kılıç Depren, S., & Depren, Ö. (2021). Housing prices in emerging countries during COVID-19: evidence from Turkey. *International Journal of Housing Markets and Analysis, ahead-of-print*(ahead-of-print). <https://doi.org/https://doi.org/10.1108/IJHMA-07-2021-0083>
- Kok, N., Koponen, E.-L., & Martínez-Barbosa, C. A. (2017). Big Data in Real Estate? From Manual Appraisal to Automated Valuation. *Journal of Portfolio Management, 43*(6), 202-211. <https://doi.org/http://dx.doi.org/10.3905/jpm.2017.43.6.202>
- Li, X., & Zhang, C. (2021). Did the COVID-19 Pandemic Crisis Affect Housing Prices Evenly in the U.S.? *Sustainability, 13*(21), 12277. <https://doi.org/https://doi.org/10.3390/su132112277>
- Neal, L., & García-Iglesias, M. C. (2013). The economy of Spain in the euro-zone before and after the crisis of 2008. *The Quarterly Review of Economics and Finance, 53*(4), 336-344. <https://doi.org/https://doi.org/10.1016/j.qref.2013.01.002>
- Nicodemo, C., & Raya, J. M. (2012). Change in the distribution of house prices across Spanish cities. *Regional Science and Urban Economics, 42*(4), 739-748. <https://doi.org/https://doi.org/10.1016/j.regsciurbeco.2012.05.003>
- Ouazad, A. (2021). Resilient Urban Housing Markets: Shocks Versus Fundamentals. In I. Linkov, J. M. Keenan, & B. D. Trump (Eds.), *COVID-19: Systemic Risk and Resilience* (pp. 299-331). Springer International Publishing. https://doi.org/https://doi.org/10.1007/978-3-030-71587-8_17
- Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications, 42*(6), 2928-2934. <https://doi.org/https://doi.org/10.1016/j.eswa.2014.11.040>
- Royo, S. (2009). After the Fiesta: The Spanish Economy Meets the Global Financial Crisis. *South European Society and Politics, 14*(1), 19-34. <https://doi.org/https://doi.org/10.1080/13608740902995828>
- Taltavull, P., & McGreal, S. (2009). Measuring price expectations. *Journal of European Real Estate Research, 2*(2), 186-209. <https://doi.org/https://doi.org/10.1108/17539260910978481>
- Wong, G. (2008). Has SARS infected the property market? Evidence from Hong Kong. *Journal of Urban Economics, 63*(1), 74-95. <https://doi.org/https://doi.org/10.1016/j.jue.2006.12.007>