

Universidad Católica de Cuyo

Facultad de Ciencias Económicas y Empresariales



Aplicaciones de Aprendizaje

Estadístico Avanzado en Economía

Franco De Giovannini

Licenciatura en Economía

Director:

Lic. Juan Marcos Pastor

Noviembre de 2021

Resumen

Aseguran algunos autores que “La información es el petróleo del siglo XXI”. El mundo de la Ciencia (o Minería) de Datos suele ser foco de noticias por los grandes avances que se han realizado en el campo de la Inteligencia Artificial. Gran parte del álgebra y de la estadística empleados para lograr dichos avances forman parte del programa de Licenciatura en Economía, por lo que implementar dichos algoritmos pueden resultar de gran utilidad para los profesionales del área. La hipótesis que propone este trabajo es que sí es necesario que los economistas conozcan las ventajas y desventajas de dichos métodos, que son moneda corriente hoy en día, para ser aplicados en un marco de un modelo económico. A efectos de exponer los pro y contras de cada método, se ajustaron modelos de regresión lineal por mínimos cuadrados ordinarios, modelos de regresión lineal con dependencia espacial por el método de máxima verosimilitud y modelos de conjunto, tanto paralelos como secuenciales. En todos los casos se utilizarán los mismos datos, los cuales fueron extraídos del sitio de clasificados Comprá en San Juan (<https://compraensanjuan.com>) en el mes de junio de 2019, utilizando técnicas de web scraping.

Introducción

“No existen soluciones, sólo trade-offs”¹. Si bien la cita está sacada completamente fuera de contexto, a la hora de resolver un problema, todos nos topamos con este tipo de decisiones. Cuando se trata de ajustar un modelo de machine learning, la disputa suele reducirse a tener un modelo para usarlo como guía o asesoramiento, o bien, a tener un modelo que prediga bien a datos nuevos. Los economistas, a lo largo de sus estudios, suelen estar expuestos primordialmente a modelos lineales, que permiten aportar interpretabilidad a la interacción entre variables endógenas y exógenas, pero, ¿Deberían los economistas utilizar técnicas de Aprendizaje Estadístico Avanzadas?

En el afán de responder la pregunta planteada, este trabajo estará compuesto por cuatro capítulos: en el primero se desarrollará el marco teórico donde se expondrá cada uno de los modelos, su álgebra e historia. Luego se hará un análisis descriptivo de los datos relevados. En el tercer capítulo se procederá a ajustar todos los modelos lineales, arrancando desde los modelos más básicos hasta los más complejos que incorporan dependencia espacial. En el capítulo final, se procederá a ajustar los modelos de conjuntos (del

¹Mitchell (1997)

inglés, ensemble methods). Finalmente, se plantean distintos escenarios en los cuales un economista podría verse beneficiado de utilizar uno u otro enfoque.

Índice general

Resumen	i
Introducción	ii
1 Aspectos metodológicos	1
1.1 Justificación	1
1.2 Marco Teórico	2
1.2.1 Origen e historia del modelo de precios hedónicos.	3
1.2.2 Ecuaciones del modelo	6
1.2.3 Aplicaciones en el mercado inmobiliario	8
1.2.3.1 Método de la variable dummy del tiempo	9
1.2.3.2 Métodos de imputación	11
1.2.3.3 Método de características	14
1.2.4 Dependencia espacial y datos geoespaciales	16
1.2.5 Métodos no paramétricos	20
1.2.5.1 Árboles de clasificación y regresión (CARTs)	21
1.2.5.2 Técnicas de conjuntos (ensemble methods)	23
1.3 Objetivos	26
1.3.1 Objetivo General	26

1.3.2	Objetivos específicos	26
1.4	Metodología	26
2	Análisis Exploratorio de Datos	28
2.1	Análisis de distribuciones	29
2.1.1	Distribuciones de clases de variables categóricas . . .	29
2.1.2	Distribuciones de variables discretas	29
2.2	Análisis y transformaciones de la(s) variable(s) objetivo . . .	32
2.3	Análisis de correlaciones	33
2.4	Ratios	33
2.5	Dependencia espacial	34
3	Especificación y estimación de modelos paramétricos	46
3.1	Especificación	46
3.1.1	Modelo Base	46
3.1.2	Modelo Base con transformación logarítmica	48
3.2	Modelos modificados	51
3.2.1	Precio por metro cuadrado (<i>pm2</i>) como variable dependiente	51
3.2.2	Logaritmo del precio por metro cuadrado <i>lpm2</i> como variable dependiente.	53
3.2.3	Estableciendo relación no lineal con variables discretas y utilizando <i>lpm2</i> como variable dependiente.	55
3.3	Modelo Spatial Lag o SAR (Simultaneous Autoregressive-Regressive)	58
3.3.1	Matriz de pesos espaciales	58
3.3.2	SAR por método de Máxima Verosimilitud	60

3.3.3	Modelo SLX	62
4	Especificación y estimación de modelos no paramétricos	66
4.1	Introducción	66
4.2	Métodos de conjuntos secuenciales	67
4.2.1	CatBoost sin variables autocorrelacionadas	67
4.2.2	CatBoost con variables autocorrelacionadas	68
4.3	Métodos de conjuntos simultáneos	69
4.3.1	Random Forest sin variables autocorrelacionadas . . .	69
4.3.2	Random Forest con variables autocorrelacionadas . .	70
	Conclusión	76
	Referencias	78

Índice de figuras

2.1	Gráfico 1	35
2.2	Gráfico 2	35
2.3	Gráfico 3	36
2.4	Gráfico 4	36
2.5	Gráfico 5	37
2.6	Gráfico 6	37
2.7	Gráfico 7	38
2.8	Gráfico 8	38
2.9	Gráfico 9	39
2.10	Gráfico 10	39
2.11	Gráfico 11	40
2.12	Gráfico 12	41
2.13	Gráfico 13	42
2.14	Gráfico 14. Mapa de dispersión de precios.	43
2.15	Gráfico 15. Grafo y polígonos de Voronoi.	44
2.16	Gráfico 16. Mapa coroplético.	45
4.1	Gráfico 17	72
4.2	Gráfico 18	72
4.3	Gráfico 19	73
4.4	Gráfico 20	73

4.5	Gráfico 21	74
4.6	Gráfico 22	74
4.7	Gráfico 22	75
4.8	Gráfico 23	75

Índice de cuadros

3.1	Modelo Base, regresión de 'precio'	49
3.2	Modelo Base, regresión de 'lprice'	52
3.3	Modelo modificado, regresión de 'pm2'	54
3.4	Modelo modificado, regresión de 'lpm2'	56
3.5	Modelo modificado, regresión de 'lpm2' y dummies discretas.	59
3.6	Modelo SAR.	62
3.7	Modelo SLX.	64

Capítulo 1

Aspectos metodológicos

1.1 Justificación

Probablemente el lector o lectora notó en los últimos años que las palabras que le recomienda el teclado de su Smartphone son cada vez más precisas y personalizadas, o quizás le pasó de estar hablando de un tema con alguien para luego revisar su teléfono y encontrar publicidad relacionada al tema. Seguramente también notó que las recomendaciones de películas o series en su cuenta de Netflix son más acertadas o que si tuvo una “conversación” con Siri o Google Assistant, estas asistentes lograron interpretar lo que les dijo verbalmente. Esto, obviamente, no es casualidad, ya que todos los ejemplos mencionados tienen algún algoritmo –o varios- de aprendizaje automático de fondo que permiten realizar predicciones sobre acciones futuras en base a características observadas de su comportamiento.

En el día a día de los economistas puede requerirse el uso de modelos econométricos, ya sea para interpretar las relaciones entre distintas

variables, así como también para realizar predicciones de alguna variable objetivo, utilizando como base algún modelo económico que permita encontrar una relación paramétrica, es decir, con relación funcional entre dichas variables. Si bien los ejemplos mencionados en el párrafo anterior no son de interés inmediato para un economista, los mecanismos utilizados para lograr los objetivos, sí. Éstos están construidos a partir de los mismos conceptos estadísticos que estudiamos, pero suelen tener una gran diferencia: no exponen una relación lineal o bien, son modelos no paramétricos.

El objetivo de este trabajo es exponer algunos métodos de aprendizaje estadístico alternativos a los vistos en el plan de estudio de la carrera, así como también los costos y beneficios de implementar una solución u otra, dadas las particularidades necesidades que puede tener una persona en el rol de economista. En efecto, la problemática que se resolverá consiste en ajustar un modelo que permita predecir el precio de un inmueble, utilizando los métodos clásicos, métodos de econometría espacial, así como también métodos de conjunto.

1.2 Marco Teórico

Como punto de partida, entendemos como *aprendizaje automático* o *estadístico –Machine Learning* en inglés– al estudio de algoritmos computacionales que mejoran automáticamente con la experiencia¹. En otras palabras, se busca desarrollar algoritmos que puedan encontrar relación y tendencias entre los datos (experiencia), es decir, aprender de los datos para

¹Mitchell (1997)

realizar alguna predicción. Difiere de la programación clásica, en la cual se deben desarrollar algoritmos manualmente a partir de los datos para obtener las respuestas buscadas. Bajo este paradigma, los modelos estadísticos como la regresión lineal o la regresión logística, forman parte de esta área de estudio, así como también el resto de algoritmos que se desarrollarán más adelante.

En cuanto al caso de estudio, es necesario ajustar un modelo econométrico que permita estimar el precio de mercado de los inmuebles radicados en la provincia. A tales efectos, el modelo implementado se denomina Modelo de Precios Hedónicos, también conocido como regresión hedónica, a través del cual se estiman los valores de las características de un bien que indirectamente afectan su precio de mercado². Por otra parte, el modelo también puede ser utilizado para estimar la demanda de un bien. Dadas estas cualidades, y su interpretación relativamente sencilla, este método ha sido ampliamente utilizado no sólo en investigaciones relacionadas al mercado inmobiliario, sino también en investigaciones sobre consumidores y mercados, para el cálculo de índices de precios al consumidor, para la valuación de autos, entre otros³.

1.2.1 ORIGEN E HISTORIA DEL MODELO DE PRECIOS HEDÓNICOS.

La creación de la metodología, según diversos autores⁴, suele ser atribuida a Andrew T. Court, quien desarrolló un índice de precios hedónicos

²Herath, Shanaka & Maier, Gunther (2010)

³En referencia a los temas mencionados, se pueden resaltar los trabajos de los autores Elizabeth C. Hirschman & Morris B. Holbrook (1982), Moulton (1996), Brian J. L. Berry & Robert S. Bednarz (1975) y Keith Cowling & John Cubbin (1972), entre muchos otros.

⁴Bartik (1987), Goodman1998 y otros.

para automóviles en 1939, en el cual utilizó el término '*hedónico*' para describir la ponderación de la importancia relativa de los varios componentes que componían a los vehículos. En su trabajo elabora problemas de no-linealidad y con cambios en los conjuntos de bienes que componen al automóvil.

Sin embargo, otros autores⁵ argumentan que el pionero en la metodología fue George C. Haas en el año 1922, aunque sin utilizar el término '*hedónico*' para describir la metodología. En el trabajo original, Haas utiliza la metodología para estimar los valores por acre ajustados por el año de venta, el tipo de carretera y el tipo de ciudad, basado en 160 transacciones realizadas en Minnesota.

El modelo eventualmente fue popularizado por Zvi Griliches a fines de los años 50s y principios de los 60s en diversos trabajos, lo que dio lugar a que los autores Lancaster (1966) y Rosen (1974) realizaran aportes –quizás los más importantes– a la base teórica del modelo. El primero estableció los fundamentos microeconómicos para analizar las características que componen los bienes y aplica dicho análisis a diversos productos. Rosen por otra parte, deriva funciones de subasta correspondientes a consumidores maximizadores de utilidad y funciones de oferta de productores maximizadores del beneficio. De esta manera, pone en evidencia un mercado implícito de características de los productos diferenciados y analiza el equilibrio del mercado. Estos precios implícitos de las características de los bienes sirven para que los productores y consumidores puedan optimizar sus elecciones de conjuntos de producción y de consumo, respectivamente.

La base teórica de Rosen deja como resultado una metodología de

⁵Peter F. Colwell & Gene Dilmore (1999) .

dos etapas para determinar los precios hedónicos: la primera consiste en estimar la ecuación hedónica; la segunda, se encarga de determinar los precios implícitos de las características a través de derivadas parciales del precio, respecto de tal característica. En este modelo, el ingreso del consumidor se toma en cuenta a la hora de definir la restricción presupuestaria, lo que implica que la intención marginal de pagar por cierto atributo puede verse modificada por un cambio en el ingreso⁶.

Las diferencias en los enfoques de Rosen y Lancaster ponen en manifiesto dos alternativas de análisis que permiten una mejor estimación dependiendo del tipo de mercado y producto. De esta forma, el análisis de Lancaster se basa en la idea que la utilidad de un bien depende de sus características y que dichos bienes pueden ser concentrados en grupos basados en ellas. La utilidad del consumidor se origina de las distintas características que los bienes poseen, no sólo en las diferentes cantidades de bienes. Esto permite que el enfoque se ajuste mejor a bienes de consumo. Finalmente, el modelo asume una relación lineal entre el precio de los bienes y las características contenidas en esos bienes.

El modelo de Rosen sostiene que existe un rango de bienes, pero que los consumidores no adquieren los atributos preferidos comprando una combinación de bienes, sino que cada bien es elegido de un espectro de ‘marcas’ y es consumido discretamente. Consecuentemente, el modelo suele ajustarse mejor a la demanda de bienes durables. En este caso, el modelo asume una relación no lineal entre el precio de los bienes y sus atributos, ya que el precio implícito no es una constante, sino una función de las cantidades del

⁶Este podría dar lugar al efecto ingreso o al efecto sustitución desarrollado en la Teoría del Consumidor.

atributo ‘comprado’ y del resto de los atributos.

1.2.2 ECUACIONES DEL MODELO

Cualquiera sea el caso del modelo especificado, el método de estimación del modelo de precios hedónicos se realiza a través del análisis de regresión múltiple, ya sea a través de Mínimos Cuadrados Ordinarios (MCO) o del método Máxima Verosimilitud. En ambos casos, se busca encontrar un vector de parámetros que mejor explique los valores de las variables explicativas. Estas variables suelen ser las características de los bienes o transformaciones matemáticas –como las dummies, o las variables instrumentales–, lo que permite contemplar relaciones de no-linealidad, interacción de variables u otras situaciones complejas. Esta información permite construir un índice de precios, el cual puede ser utilizado para comparar distintas ciudades (sección cruzada) o una misma ciudad en el tiempo (serie de tiempo).

La forma estándar de estimación de la regresión toma la forma:

$$P = f(I, N, L, C, t)$$

donde P es el precio del inmueble (o el valor del alquiler), I son los atributos relativos al inmueble, N son las características del barrio, L son variables particulares del mercado, C son las condiciones contractuales y t es un indicador del tiempo.

Dada la complejidad en la forma de medición de algunas variables y la disponibilidad de datos relacionados, es poco común ver todas las variables utilizadas en un mismo trabajo. Según Malpezzi (2003), las variables que

habitualmente se utilizan son:

- Número y tipo de ambientes (baños, dormitorios, etc.)
- Tipo de inmueble (departamento, casa, casa quinta, terreno, etc.)
- Antigüedad
- Tipo de construcción (en seco, Steel frame, tradicional)
- Características estructurales (sótano, cochera, chimenea, piscina, etc.)
- Servicios disponibles (gas, electricidad, agua, cloacas)
- Instalaciones climáticas (calefacción, piso radiante, A/C, etc.)
- Superficie del terreno
- Superficie construida

En cuanto a la forma funcional, la estimación de la regresión de la ecuación puede ser lineal, semi logarítmica o log-log, siendo la segunda la más utilizada. Este método presenta la ventaja de que los coeficientes de las variables estimadas representan las proporciones del precio que son directamente atribuibles a las características. Por su parte, la implementación del método log-log permite estimar las elasticidades de cada una de las características de los inmuebles.

Con el desarrollo del modelo y su aplicación al mercado inmobiliario, distintos autores⁷ han incluido variables que toman en cuenta la dimensión espacial en presencia de dependencia espacial y autocorrelación espacial. A tales efectos, una de las formas más sencillas de incluir aspectos espaciales en el modelo es incluir una variable de distancia entre la propiedad y el distrito comercial. Sin embargo, el análisis espacial suele ser considerado un arma de

⁷Se pueden resaltar: Jackson (1979), Dubin (1988), Anselin (1988), Basu, Sabyasachi & Thibodeau, Thomas G (1998), entre otros.

doble filo: permite obtener resultados más robustos y explicativos, pero al mismo tiempo, complejiza el análisis considerablemente.

1.2.3 APPLICACIONES EN EL MERCADO INMOBILIARIO

El modelo de precios hedónicos suele ser una herramienta conveniente en un contexto de estimación de precios de viviendas en distintas situaciones, tales como la generación de un índice de precios ajustado por la calidad de la construcción, la generación de una herramienta de tasación automática o masiva y para explicar las variaciones de precios o determinar el impacto en los precios de las viviendas a través de ciertas características.

Dependiendo de las necesidades, los objetivos del modelo y la disponibilidad de datos, existen diversas formas de encarar la estimación. Una primera consideración es si se utilizan datos de operaciones realizadas o si se realiza en base al stock de inmuebles disponibles. La primera tiene la ventaja que, al tratarse de operaciones efectivizadas, se tiene certeza de los valores a los cuales se concretaron dichas transacciones. Además, dicha información suele ser aportada por instituciones colegiadas de profesionales o cámaras empresarias del sector, por lo cual suele haber más datos de características individuales disponibles, pero una cantidad menor de observaciones, esparcidas en el tiempo. Si se trata de inmuebles que forman parte del stock disponible, suelen haber más observaciones, pero con menos datos individuales y, por otra parte, no se tiene certeza del precio. Esto se debe a que no se tiene noción si los inmuebles publicados ya fueron comercializados o no, ni el precio real al cual se concretó la operación. Además, aparece un riesgo de sesgo de selección de observaciones, dado que existe una correlación entre el

precio del inmueble y los datos disponibles sobre sus características: los inmuebles de mayor valor suelen disponer de más detalles de sus características que los de menor valor.

Ya sea que se cuente con datos de transacciones efectivizadas o con el stock de inmuebles, existen tres métodos de estimación del modelo, cada uno con sus ventajas y sus desventajas, que permiten que sean utilizados en distintas ocasiones con mayor grado de éxito, las cuales serán explicadas a continuación.

1.2.3.1 Método de la variable dummy del tiempo

Es el método estándar de estimación del modelo de precios hedónicos y habitualmente utiliza la forma semilogarítmica para estimar el peso relativo de cada variable explicativa. Dicha metodología consiste en generar una variable dummy que capte el tiempo de acuerdo a la fecha en la que se realizó la transacción, a efectos de actualizar los valores históricos de las transacciones realizadas. La forma funcional es la siguiente:

$$y = Z\beta + D\delta + \varepsilon$$

donde y es un vector de dimensiones $H \times 1$ de elementos $y = \ln p_h$ siendo p_h el precio de cada inmueble, Z es una matriz $H \times C$ de características del inmueble, β es un vector $C \times 1$ de los precios sombra de las características, D es una matriz $H \times T - 1$ de variables dummies de períodos, δ es un vector $T - 1 \times 1$ de precios de períodos y ε el vector de errores aleatorios. H , C y T representan el número de inmuebles, el número de características y los

períodos de tiempo del conjunto de datos. El formato permite incorporar funciones e interacciones de características, entre las cuales se destacan el cuadrado de los metros cuadrados o el producto entre dormitorios y baños.

Si el objetivo es construir un índice de precios ajustado por la calidad del inmueble, entonces el interés principal recae en los parámetros que miden los efectos fijos en los logaritmos de los niveles de precios luego de controlar los efectos de las diferencias en los atributos de los inmuebles. De esta manera, el índice de precios para el período t se obtiene simplemente exponiendo el valor estimado de δ_t , tal que:

$$\hat{P}_t = \exp[\hat{\delta}_t + \frac{1}{2}\hat{\sigma}_t^2]$$

Por otra parte, debido a que uno de los determinantes clave del precio de un inmueble es la ubicación, las estimaciones se pueden mejorar considerablemente si se la incluye. La forma que solía utilizarse en el pasado era incorporar una dummy basada en el código postal del inmueble. Actualmente, debido a la disponibilidad de datos de geolocalización y su fácil aplicación, se puede modelar la dependencia espacial lo que permite resultados más robustos y explicativos.

Este método tiene como principal ventaja su simplicidad, dado que permite estimar la variación de precios en el tiempo de manera sencilla e intuitiva, utilizando la forma funcional semilogarítmica. Por supuesto, no está exento de limitaciones, entre las que se pueden destacar el hecho de que la inclusión de un año extra de observaciones afecta el índice de los períodos anteriores o la falta de flexibilidad del modelo. Una forma de aumentar la

flexibilidad consiste en interactuar las variables características del inmueble con las variables del tiempo, ya que la demanda por ciertas particularidades puede variar entre períodos. La inclusión de tales interacciones, por su parte, reduce los grados de libertad de la estimación.

Debido a que la inclusión de datos de nuevos períodos afecta el índice de los períodos anteriores, este método ha sido implementado con poca popularidad en un contexto de estimación de precios de viviendas. Lo mencionado anteriormente implica una violación de la fijación temporal, lo que significa que todos los resultados tienen que volver a ser computados cuando un período nuevo es agregado a la base de datos.

1.2.3.2 Métodos de imputación

Los métodos de imputación utilizan fórmulas estándares de índices de precios, como los de Laspeyres y Paasche, que son ampliamente utilizados en la construcción de índices de precios al consumidor. Ambas fórmulas miden el cambio en el precio de un conjunto dado de bienes en el tiempo. Esto requiere que el precio de cada bien en la canasta se encuentre disponible para cada período, lo cual no ocurre habitualmente en el mercado inmobiliario. Esto se debe a que es muy poco probable que una misma parcela se encuentre a la venta y se negocie en dos períodos distintos, por lo tanto, no es posible computar los índices de precios de Laspeyres o Paasche basados en transacciones reales. Ante tal impedimento, se recurre a la *imputación* del precio de dicho bien en los períodos en los cuales no se vendió el mismo, lo cual se realiza estimando el precio \hat{P}_{th} en base a los estimadores obtenidos con la regresión hedónica.

La utilización de este método agrega una nueva dimensión al problema del índice de precios, dado que otorga cierta discreción a quien construye el índice. Si el precio de una casa h no se encuentra disponible en el período t , no queda otra que imputar el precio estimado; pero si el precio sí se encuentra disponible para tal período, existe aún la tentación a imputar el precio ya que permite reducir el sesgo de variables omitidas.

Para poder implementar el método, definimos $\hat{p}_{th}(z_{sh})$ como el precio estimado en el período t de un inmueble vendido en el período s , lo cual se realiza reemplazando las características de la casa h vendida en el período s dentro de los parámetros estimados del modelo hedónico del período t , tal que:

$$\hat{p}_{th}(z_{sh}) = \exp\left(\sum_{c=1}^C \beta_{ct} \hat{Z}_{csh}\right)$$

A dicha estimación corresponde corregir el sesgo⁸, lo cual realizamos agregando la mitad de la varianza del error de la ecuación de la regresión hedónica, lo que nos deja:

$$\hat{p}_{th}(z_{sh}) = \exp\left(\sum_{c=1}^C \beta_{ct} \hat{Z}_{csh} + \frac{1}{2} \sigma_t^2\right)$$

Con el resultado obtenido, se procede a la estimación de la fórmula de índice de precios, en este caso utilizaremos Laspeyres, que se puede definir como:

$$P_{st}^L = \sum_{h=1}^{H_s} \left\{ w_{sh} \left[\frac{\hat{p}_{th}}{p_{sh}} \right] \right\} = \frac{\sum_{h=1}^{H_s} p_{th}}{\sum_{h=1}^{H_s} p_{sh}}$$

donde w_{th} son las ponderaciones de observaciones imputadas definidas

⁸Esto se debe a que estamos estimando $p_{th}(z_{sh})$ y no $\ln p_{th}(z_{sh})$.

por:

$$w_{th} = \frac{p_{th}}{\sum_{m=1}^{H_t} p_{tm}}$$

Ahora simplemente resta incorporar $\hat{p}_{th}(z_{sh})$ en la fórmula de P_{st}^L , lo que resulta en:

$$L_1 : P_{st}^{L_1} = \sum_{h=1}^{H_s} \left\{ w_{sh} \left[\frac{\hat{p}_{th}(z_{sh})}{p_{sh}} \right] \right\} = \frac{\sum_{h=1}^{H_s} \hat{p}_{th}(z_{sh})}{\sum_{h=1}^{H_s} p_{sh}}$$

$$L_2 : P_{st}^{L_1} = \sum_{h=1}^{H_s} w_{sh} \left[\frac{p_{th}}{s_{th}} \right] + \sum_{h=1}^{H_s} \left\{ w_{sh} \left[\frac{\hat{p}_{th}(z_{sh})}{p_{sh}} \right] \right\} = \frac{\left[\sum_{h=1}^{H_s} p_{th} + \sum_{h=H_{st}+1}^{H_s} \hat{p}_{th}(z_{sh}) \right]}{\sum_{h=1}^{H_s} p_{sh}}$$

L_1 y L_2 representan dos formas distintas de estimar el índice, con resultados distintos. La diferencia surge en las observaciones que se toman: L_1 imputa todos los precios al período t , mientras que L_2 sólo imputa los precios de los inmuebles que no estaban disponibles en el período s y t . Existen dos alternativas más: una de ellas consiste en imputar todos los precios para los períodos s y t ; la otra sería imputar todos los precios para ambos períodos excepto los de los inmuebles que tengan ambos datos⁹. Si se utiliza cualquiera de los últimos dos métodos mencionados, se recurre al caso de *doble imputación*, dado que en ambos se reemplazan valores reales de los períodos s y t por valores estimados. Esta metodología, si bien parece contradictoria o no deseada, puede ser útil en ciertas ocasiones ya que permite reducir el sesgo de variables omitidas. Supongamos que hay una observación la cual tiene precio estimado mayor al real, tal que: $\hat{p}_{th}(z_{sh}) > p_{sh}$

⁹Para ver una representación algebraica de dichas alternativas, ver Hill (2011)

Esto puede ser porque el comprador obtuvo una ganga o porque el inmueble tiene malos indicadores de variables omitidas (como puede ser polución o crimen si no estuvieran contempladas en el modelo original). Si fuera cierto el segundo caso, $\hat{p}_{th}(z_s h)$ va a sobreestimar el precio real del inmueble, ya que $\hat{p}_{th}(z_s h)/p_{sh}$ va a tener un sesgo positivo. Por el contrario, si utilizamos las características del inmueble para estimar el precio en el período s y en el período t , la variación del precio en el tiempo de dicha observación no estará sesgada, ya que ambas estimaciones tendrán incorporadas las variables omitidas.

Esta metodología de estimación del modelo es un tanto más flexible que el basado en dummies de tiempo, ya que permite imputar datos en períodos en los que no se disponen los mismos, aumentando la cantidad total de observaciones. Por supuesto que la metodología no está libre de críticas: por un lado, otorga demasiada discrecionalidad a quien construye el índice, por otra parte, requiere la estimación del modelo por cada período individualmente, lo que no permite realizar interacciones intertemporales entre las variables.

1.2.3.3 Método de características

Del mismo modo que el método de imputación, esta metodología requiere calcular el modelo hedónico para cada período por separado y requiere del uso de fórmulas de índices de precio estándares. La particularidad de esta metodología es que se construye un inmueble “tipo”, en base a los promedios, para cada período y luego se le imputan los precios de cada período, como función de sus características utilizando los precios sombras derivados del

modelo hedónico. El inmueble “tipo” o “promedio” puede ser la media o la mediana. La diferencia de precios se obtiene, consecuentemente, dividiendo el precio imputado del inmueble para los dos períodos, lo que lo convierte en un modelo de doble imputación.

Cabe destacar que el inmueble tipo o promedio, puede ser tomado en cualquiera de los dos períodos de análisis y no requiere que sea construido con variables discretas, es decir, si tomamos la cantidad promedio de baños, dormitorios, metros cuadrados, metros construidos, etc., podemos utilizar unidades no enteras (p. ej.: 2,42 dormitorios, 1.65 baños, etc.). Una vez calculado el precio estimado para el inmueble de esas características en cada período, se construye el índice de precios utilizando nuevamente Laspeyres o Paasche, como se demuestra a continuación:

$$L : P_{st}^L = \frac{\hat{p}_t(\bar{z}_s)}{\hat{p}_s(\bar{z}_s)} = \exp \left[\sum_{c=1}^C (\hat{\beta}_{ct} - \hat{\beta}_{cs}) \bar{z}_{cs} \right]$$

$$\text{con } \bar{z}_{cs} = \frac{1}{H_s} \sum_{h=1}^{H_s} z_{csh}$$

$$P : P_{st}^P = \tilde{P}_{st}^P = \frac{\hat{p}_t(\bar{z}_s)}{\hat{p}_s(\bar{z}_s)} = \exp \left[\sum_{c=1}^C (\hat{\beta}_{ct} - \hat{\beta}_{cs}) \bar{z}_{ct} \right]$$

$$\text{con } \bar{z}_{ct} = \frac{1}{H_s} \sum_{h=1}^{H_s} z_{cth}$$

donde H_s y H_t son el número total de inmuebles observados en cada período, \bar{z}_{cs} y \bar{z}_{ct} son los vectores de características de los inmuebles promedio de cada período. Para extender el análisis se puede calcular la media geométrica de ambos índices y obtener un híbrido de doble imputación, que

es análogo al índice de Fisher, tal que:

$$F : P_{st}^F = \sqrt{P_{st}^L \times P_{st}^P} = \exp \left[\frac{1}{2} \sum_{c=1}^C (\hat{\beta}_{ct} - \hat{\beta}_{cs})(\bar{z}_{cs} + \bar{z}_{ct}) \right]$$

Debido a este método realiza imputación de valores, sufre de los mismos defectos y virtudes que aquél. Otro de los inconvenientes que se presenta con el uso de esta metodología es cuando se incluyen ajustes por dependencia espacial. Esto se debe a una sencilla razón: ¿cómo determinamos la “*ubicación promedio*” de los inmuebles? A pesar de esto y debido a la sencillez instrumental del método, éste es el más utilizado para computar índices de precios. Otra ventaja que es menester señalar de la metodología es su utilidad para comparar índices de precios entre diferentes regiones.

1.2.4 DEPENDENCIA ESPACIAL Y DATOS GEOESPACIALES

En la industria inmobiliaria –principalmente en países angloparlantes– existe una frase un tanto cómica que resume las tres reglas del real estate: “ubicación, ubicación y ubicación”. Como uno puede imaginarse, dicha variable es una de las más importantes a la hora de determinar el precio de un inmueble. Debido a este motivo, es lógico suponer que existe dependencia espacial en los precios de los inmuebles¹⁰, ya que muchos de los determinantes del precio de los inmuebles suelen ser compartidos por las propiedades vecinas. Las parcelas de los barrios suelen ser edificadas simultáneamente, lo que resulta en diseños y características similares; inclusive las propiedades vecinas disfrutan de las mismas comodidades locales.

¹⁰Pace, R Kelley & Gilley, Otis W (1997), Basu, Sabyasachi & Thibodeau, Thomas G (1998), Bourassa, Steven et al. (2002) son algunos ejemplos de estudios empíricos que alertan la existencia de dependencia espacial en los residuos.

En econometría es habitual suponer la independencia de las observaciones para poder realizar un análisis de inferencia estadístico. Una de las formas de explicitar matemáticamente esto es suponer que los valores observados y_i , para un individuo i , son estadísticamente independientes de otros valores y_j , para el individuo j , tal que $E(y_i y_j) = E(y_i)E(y_j) = 0$, donde $E(\cdot)$ es el operador de la esperanza. La existencia de correlación espacial implica una violación los supuestos de Gauss-Márkov¹¹, en este caso $E(y_i y_j) \neq 0$, lo que nos da como resultados estimadores inefficientes – en el mejor de los casos- o estimadores sesgados, ya que los regresores pueden estar correlacionados entre ellos o con respecto al residuo. Por lo tanto, la inclusión explícita de la dependencia espacial ayuda a eliminar el problema de variables locales omitidas.

Existen dos formas de incluir la variable *ubicación* en el modelo: la primera es generando una dummy de cada barrio, lo cual requiere tener información precisa de los límites de los mismos y la construcción de tantas dummies como barrios se contemplen; la segunda es a través de las coordenadas de latitud y longitud. Cualquiera sea el caso, se requiere la creación de una matriz W de ponderaciones espaciales, la cual tendrá dimensiones $H \times H$ donde H es la cantidad de observaciones incluidas en la regresión. A efectos de simplificar los valores y el peso computacional, se suele utilizar el algoritmo de triangulación de Delaunay, que simplifica la matriz a ceros y unos. Dicho algoritmo utiliza los datos de longitud y latitud de cada inmueble y crea un conjunto de triángulos en los que no existen más puntos que los vértices del mismo en su circunferencia circunscrita. Dos inmuebles se consideran vecinos si ambos son vértices del mismo triángulo y toman valor

¹¹Principalmente las condiciones de no colinealidad, exogeneidad y homocedasticidad.

1 en la matriz, esto es, si los inmuebles j y k son vecinos, entonces las posiciones jk y kj toman el valor 1 en la matriz, de lo contrario toman el valor cero. Además, la diagonal de ubicaciones jj y kk toman valor 0 por defecto. Un atajo que se realiza en la práctica, tiene que ver con la creación de los polígonos de Voronoi a partir de los puntos observados. Con estos polígonos se procede a considerar vecinos en base a un criterio de contigüidad como Rook, Bishop o Queen, que responden al tipo de contigüidad que existen en las casillas de ajedrez, en base al tipo de movimiento de dichas piezas (Torre, Alfil o Reina).

Una vez creada la matriz de ponderaciones espaciales W , la dependencia espacial puede ser capturada en un modelo hedónico simultáneamente autorregresivo (SAR), que puede ser definido de la siguiente forma:

$$y = \beta X + u.u = \lambda W u + \varepsilon.$$

donde es un vector $H \times 1$ de errores aleatorios. Como siempre, se estima el valor del vector de dimensiones $C \times 1$ y, en este caso, se estima el escalar λ , el cual mide la influencia porcentual local promedio de las observaciones vecinas para cada observación. Es decir, mide el porcentaje de la variación en u_h que es causado por las influencias locales de los vecinos de h . Lo atractivo del modelo es que sólo requiere construir W y estimar los parámetros β y λ , lo cual se realiza utilizando el método de máxima verosimilitud –en vez de Mínimos Cuadrados Ordinarios–.

Existen otras alternativas al modelo SAR, como por ejemplo la creada por Lesage (1999) y denominada “modelo simultáneo autoregresivo-regresivo”, que captura la dependencia espacial de la siguiente forma:

$$y = \rho W y + X\beta + \varepsilon$$

donde los parámetros a estimar son ρ y β con el método de máxima verosimilitud, nuevamente.

Una alternativa al método de triangulación de Delaunay es utilizar el de los *k vecinos más cercanos*, el cual consiste en tomar una observación y seleccionar la cantidad k de vecinos en el espacio que se encuentren a menor distancia. Este enfoque fija siempre la cantidad k de vecinos, mientras que Delaunay sólo considera vecinos a los que se encuentran dentro del área de circunferencia circumscripita, por lo que cada observación puede obtener una cantidad distinta de vecinos. Aquí yace la principal desventaja del método de los *k vecinos más cercanos*, donde al forzar la cantidad de vecinos, los *outliers* -observaciones atípicas- tienen una ponderación mayor de la esperada. Aun así, es un método de fácil estimación que produce estimadores considerablemente mejores que si no se tuviera en cuenta la dimensión geoespacial. Habitualmente se lo suele tomar como punto de partida para conocer los efectos de los atributos de las observaciones y luego ajustar su eficacia con métodos más robustos.

Si bien existen múltiples métodos que buscan interpretar la dependencia espacial que tienen los precios de los inmuebles, éstos no suelen ser utilizados en índices de precios, ya que la mayoría de los índices suelen ser construidos a través del método de características. Recordemos que esta metodología crea un inmueble promedio y estima la variación del precio en el tiempo, pero cuando se trata de ubicación, el concepto de promedio no tiene mucho sentido: se generaría un punto que poco tiene que ver con el precio

promedio de la zona donde caiga; más aún, podría caer en el agua, si se tratara de una ciudad desarrollada sobre un puerto, una bahía o sobre ambos lados de un río.

1.2.5 MÉTODOS NO PARAMÉTRICOS

Los métodos no paramétricos se caracterizan por no asumir ninguna relación funcional del modelo hedónico, por lo que evitan cualquier problema de especificación errónea, es decir, no suponen una relación lineal o cuadrática o de cualquier tipo entre las variables descriptivas y las dependientes. Esto las convierte en una gran herramienta para grandes conjuntos de datos, ya que en tales situaciones obtienen menores errores para predicciones fuera de la muestra. Existe un universo de modelos no paramétricos que han tomado relevancia en los últimos años con el *boom* de la disponibilidad de datos de gran escala –*Big Data*– y el desarrollo de la disciplina de Ciencia de Datos. Entre ellos se destacan los árboles de decisión, *Random Forest*, *Gradient Boosting* y las *Redes Neuronales*, estas últimas conocidas como *Deep Learning*.

Al no tener una forma funcional establecida, estos métodos gozan de una mayor flexibilidad para realizar las estimaciones. Al mismo tiempo, dicha particularidad no les permite estimar los precios sombras de las características. Por este motivo es poco común encontrarse con índices de precios de viviendas construidos a partir de esta metodología, ya que el objetivo principal de los índices de precios es estimar la variación de los mismos en el tiempo, sean los precios de los inmuebles o los precios sombra de las características. Debido a esto, esta metodología suele utilizarse para construir

índices de precios espaciales, en vez de índices de precios temporales, lo que lo convierte en una gran herramienta para comparar precios entre regiones en un mismo periodo, así como también para predecir el precio de los inmuebles, habitualmente visto en herramientas de tasación online.

1.2.5.1 Árboles de clasificación y regresión (CARTs)

Los árboles de decisión arrancan desde un paradigma de las ciencias de la computación bastante sencillo, a partir del cual realizan particiones binarias en el espacio multidimensional de las variables exógenas. En otras palabras, van generando rectángulos en el espacio multidimensional donde se encuentran los datos y ajustan un modelo sencillo en cada uno de estos, tal como una constante. En el caso de la regresión, a cada subespacio generado le asignan el valor medio de los datos. Cada partición *-nodo-* tiene una descripción sencilla como $X_1 = t_1$ a partir de la cual el algoritmo genera dos regiones *-ramas-* R_1 y R_2 , una para valores menores o iguales a t_1 y otra para los valores mayores. De esta manera, el algoritmo predice la constante c_1 para la región R_1 y c_2 para la región R_2 . En el caso de una regresión, los valores de c_i están dados por la media para la región R_i .

Formalmente, considerando un conjunto de datos con p variables independientes y N observaciones, es decir (x_i, y_i) con $i = 1, 2, \dots, N$ y $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$, se generan M particiones R_1, R_2, \dots, R_M y se asignan las constantes c_m para cada región, tal que:

$$\hat{f}(X) = \sum_{m=1}^M c_m I(x_i \in R_m)$$

Utilizando el criterio de minimización de la suma de errores cuadráticos $\sum(y_i - f(x_i))^2$, la mejor estimación de \hat{c}_m viene dada por el promedio y_i en la región de R_m :

$$\hat{c}_m = \text{ave}(y_i | x_i \in R_m),$$

donde ave es el promedio de y_i para las x_i pertenecientes a R_m .

Por supuesto, la clave del algoritmo está en encontrar los puntos donde realizar los cortes de las p variables. Para conseguirlo, se busca minimizar la suma de los errores cuadráticos de ambas regiones generadas por un corte en x . Matemáticamente se puede definir un corte s en la variable j , que genera dos regiones:

$$R_1(j, s) = \{X | X_j \leq s\} \text{ y } R_2(j, s) = \{X | X_j > s\}.$$

Para encontrar s , se busca el valor que resuelve:

$$\min j, s \left[\min_{c_1} \sum_{x_i \in R_1(j, s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j, s)} (y_i - c_2)^2 \right]$$

donde para cada par (j, s) las minimizaciones de c_1 y c_2 son resueltas por:

$$\hat{c}_1 = \text{ave}(y_i | x_i \in R_1(j, s)) \text{ y } \hat{c}_2 = \text{ave}(y_i | x_i \in R_2(j, s)).$$

El problema inherente de este algoritmo, es que el árbol de regresión puede crecer hasta el punto en el cual cada observación coincide con la cantidad de *hojas* o nodos terminales, lo cual genera un sobreajuste a los datos de entrenamiento y pérdida de generalización del modelo a datos no observados. Esto se debe a que si el modelo tiene i hojas en un conjunto de i observaciones, el error cuadrático del modelo, será 0, ya que éste podrá predecir siempre el mismo valor con el que fue entrenado, generando un $R_2 = 1$.

Hasta dónde dejar crecer el árbol es cuestión de cada problemática específica que se esté analizando. Existen dos formas de evitar el sobreajuste: una consiste en establecer una condición para la generación de *ramas* o nodos, mientras que la otra se conoce como *poda*. La primera es un tanto arriesgada, ya que, si existen ramas valiosas en otra variable, pero no se cumple con la condición establecida para realizar la partición, el algoritmo no podrá encontrar dicha rama. La segunda consiste en permitir el sobreajuste y luego recortar las hojas y ramas que provocan el sobreajuste a los datos de entrenamiento.

Una alternativa usual, consiste en encontrar la cantidad de ramas y hojas del árbol de regresión generando n cantidad de modelos, de complejidad variada y evaluando la capacidad de generalización del mismo utilizando *validación cruzada*.

Esta técnica consiste en subdividir los datos de entrenamiento en distintos grupos con los cuales se entrenará y evaluará el comportamiento del modelo. Es una técnica habitualmente utilizada en la ciencia de datos con el objetivo de encontrar los *hiper-parámetros* que optimizan el modelo. Es decir, aquellos parámetros del modelo -por ejemplo, la cantidad de hojas y ramas de un árbol de regresión- que debe definir el investigador a su criterio y que no se pueden definir con las funciones de costo asociadas al modelo.

1.2.5.2 Técnicas de conjuntos (ensemble methods)

Las técnicas de conjunto consisten en ajustar distintos modelos en virtud de minimizar el error de estimación y combatir la elevada varianza de

los modelos complejos, es decir, el sobreajuste. Éstas se pueden agrupar en dos categorías:

- Técnicas de conjunto simultáneas o paralelas.
- Técnicas de conjunto secuenciales o aditivas.

La primera consiste en ajustar una gran cantidad de modelos complejos (de elevada varianza, es decir, que sobreajustan a los datos) y promediar los resultados de dichos modelos. El mecanismo utilizado para generar resultados robustos se conoce como *Bagging* o **Bootstrap Aggregating** y su funcionamiento es bastante sencillo: dado un set de datos $Z = (z_1, z_2, \dots, z_N)$ donde $z_i = (x_i, y_i)$, se realiza una selección aleatoria de N observaciones z_i con reposición. De esta manera, obtendremos una cantidad B de conjuntos de datos creados a partir del set original, donde cada uno de ellos tendrá N cantidad de observaciones. Esta metodología se conoce como *Bootstrap* y permite aumentar sintéticamente el tamaño de los conjuntos de datos. Una vez generados los B conjuntos de datos, se realiza un ajuste del modelo complejo a cada uno de estos conjuntos. Finalmente, los resultados de éstos se promedian para poder estimar las medidas de error totales -o el R^2 .

Una aplicación especial de esta metodología se conoce como *Random Forest*, donde además de lo descrito, cada árbol de decisión sólo puede elegir el mejor nodo de un subconjunto de los vectores de características, elegido aleatoriamente. Cabe destacar que todos los algoritmos que pertenecen a esta familia de conjuntos, buscan que cada uno de los árboles que lo componen sean lo más diferentes entre sí, a efectos de que el promedio de los valores estimados por cada uno de ellos, se aproxime al valor observado de y_i .

La segunda familia de algoritmos de conjuntos, conocida como *Gradient Boosting*, parte de un modelo sencillo y sesgado $f_1(X)$ para posteriormente ajustar otro modelo $f_2(X)$ ponderando las observaciones en base a los errores de estimación de cada una de ellas, de acuerdo a una función de pérdida definida. Existen varias implementaciones entre las que se destacan *AdaBoost*, *XGBoost*, *LightGBM* y *CatBoost*, que principalmente varían sobre la metodología que utilizan para ponderar los errores de las observaciones, así como también en los mecanismos de uso de recursos computacionales. Genéricamente, se los puede definir como:

$$G(X) = \text{sign}\left(\sum_{m=1}^M \alpha_m G_m(x)\right)$$

donde $\alpha_1, \alpha_2, \dots, \alpha_M$ son los pesos estimados por el algoritmo y ponderan la contribución del modelo $G_m(X)$ a la estimación final. Este mecanismo se conoce como “comité”, ya que cada modelo $G_m(X)$ tendrá un voto asociado a su relevancia, de acuerdo a su capacidad de minimizar el error de estimación. Para calcular los α_m del regresor final, cada paso secuencial del algoritmo genera un peso w_1, w_2, \dots, w_N a cada una de las observaciones (x_i, y_i) en base al error de estimación de y_i . Inicialmente cada observación tiene un peso tal que $w_i = \frac{1}{N}$ y luego se van ponderando las observaciones con mayor error de la manera previamente descrita¹².

¹²Para ver las funciones de pérdida de AdaBoost y sus particularidades, ver Trevor Hastie et al. (2009), capítulo 10.

1.3 Objetivos

1.3.1 OBJETIVO GENERAL

Analizar y modelar un mismo problema económico desde distintos enfoques estadísticos y de aprendizaje automático a efectos de representar las ventajas y desventajas de cada una de ellas.

1.3.2 OBJETIVOS EXPECÍFICOS

1. Desarrollar un marco conceptual para interpretar las implicancias matemáticas y estadísticas de cada enfoque.
2. Realizar un análisis descriptivo del conjunto de datos para exponer el problema en cuestión.
3. Ajustar distintos modelos de regresión lineal multivariada a los datos.
4. Desarrollar distintos modelos de dependencia espacial para analizar la aleatoriedad de la distribución de nuestra variable objetivo.
5. Entrenar modelos de machine learning con el conjunto de datos.
6. Comparar los resultados.

1.4 Metodología

Debido a la accesibilidad y disponibilidad de datos, las estimaciones sobre los precios de los inmuebles se realizarán sobre el stock de inmuebles del mercado, es decir, sobre la oferta actual que existe en la provincia, en vez de las transacciones realizadas. Las distintas instituciones referentes de

la industria son un tanto discretas a la hora de compartir información por una sencilla razón: existe una gran disparidad de valores entre el avalúo fiscal y el precio real de mercado, por lo que no quieren dejar expuestos a los clientes que realizaron una transacción inmobiliaria por un monto artificialmente bajo. A la hora de elegir entre transacciones realizadas y stock inmobiliario existe un *trade-off* inevitable: certeza de precios vs. representatividad de la oferta inmobiliaria. Las transacciones efectivas proveen certeza del precio por el cual se vendieron los inmuebles, pero pueden ser no representativas del total de inmuebles que se comercializan. Esto puede surgir en situaciones en las que operaciones de ciertos tipos de inmuebles o de ciertos montos no se realicen a través de inmobiliarias, por ejemplo. En cuanto a la base de datos que se utilizará para estimar el modelo, la misma parte de un relevamiento de inmuebles disponibles a la venta en distintas plataformas de clasificados online que se utilizan en la provincia. Respecto del método de estimación del modelo, se utilizará el de características, acompañado del modelo SAR ya que en un primer momento nos interesa conocer los precios por zonas y tipos de inmuebles. Finalmente se mostrará el uso de modelos no paramétricos de conjunto, particularmente *Random Forest* y *Gradient-Boosting* en su implementación a través de *CatBoost*. La construcción del índice de precios servirá para ajustar dichos valores en el tiempo. La especificación de los modelos paramétricos se realizará en el Capítulo III, mientras que en el Capítulo IV se encontrará la de los métodos de conjuntos.

Capítulo 2

Análisis Exploratorio de Datos

Con la intención de conocer los datos sobre los cuales se ajustarán los modelos, se procede a realizar un análisis en detalle de cada una de las dimensiones del dataset. Las mismas son:

- Price
- Price US
- LotLandSize
- ConstructionSize
- Bedrooms
- Bathrooms
- CarParkingLots
- District
- Latitude
- Longitude
- PropertyType

y contiene 731 observaciones en total. A efectos de homogeneizar la unidad de medida de los valores de los inmuebles, todos los inmuebles se encuentran valuados en pesos a un tipo de cambio de 43 ARS por 1 USD, correspondiente al día 1 de Junio de 2019.

2.1 Análisis de distribuciones

2.1.1 DISTRIBUCIONES DE CLASES DE VARIABLES CATEGÓRICAS

En el gráfico 2.1 se puede apreciar la cantidad de avisos según tipo de inmueble. El tipo más popular corresponde a Lotes, con 438 unidades (60% de las observaciones), seguido de Casas con 194 filas (25,5% de las observaciones) y finalmente los Departamentos con 99 observaciones (14,5%). Con respecto a la cantidad de avisos según la localidad del inmueble, en el gráfico 2.2 se puede ver que tanto Capital como Santa Lucía tienen 179 observaciones (24,5%), seguido de Rivadavia con 159 (21,8%), Rawson con 126 (17,2%), Chibas con 61 (8,3%) y finalmente Pocito con 27 observaciones, es decir, el 3.7%.

2.1.2 DISTRIBUCIONES DE VARIABLES DISCRETAS

Con respecto a la cantidad de baños, el valor máximo observado fue de 5 unidades, que ocurrió en una única observación (0.13%). De la misma forma, 4 baños fue observado sólo en 4 inmuebles, es decir 0.54% de las propiedades. Por el contrario, debido a que la mayoría de los inmuebles son

lotes, el 438 (60%) inmuebles no poseen baños, mientras que los inmuebles que sí poseen, muestran una relación decreciente:

- 25% ó 185 observaciones tienen sólo 1 baño,
- 11% ó 81 observaciones tienen 2 baños y,
- 3% ó 22 observaciones tienen 3 baños.

Para analizar más de cerca la distribución de los baños, se realizó una tabla dinámica donde se agrupó las observaciones por tipo y cantidad de baños para contar la cantidad de observaciones en cada par de valores. El resultado se puede observar en el gráfico 2.4, donde queda expreso que del total de inmuebles que poseen un baño, 50,8% corresponden a Casas (94 unidades) y 49,2% (91 observaciones) corresponden a Departamentos. La diferencia se ve ampliada cuando nos enfocamos en inmuebles con 2 baños, donde el 92,5% de los mismos son Casas, es decir, 75 inmuebles, mientras que el resto (6 observaciones) corresponden a Departamentos. Finalmente, si bien no se observaron departamentos con 3 baños, sí se relevaron 2 departamentos con 4 baños, misma cantidad que casas.

El histograma de los dormitorios se puede observar en el gráfico 2.5. En el mismo se utilizó el tipo de inmueble para resaltar la distribución entre los mismos. Dejando de lado los Lotes por razones obvias, se puede apreciar que en departamentos, la cantidad más frecuente es un dormitorio (42 observaciones), mientras que las casas frecuentan tener tres unidades (102 observaciones). De los inmuebles con un dormitorio, el 89% fueron departamentos, en cuanto a los inmuebles con dos dormitorios, la mayoría son casas (44 unidades, 70%) mientras que sólo se observaron 19 departamentos (30%). Por otra parte, se observó que el 85% de los inmuebles con 4 dormitorios son

casas (29 observaciones), mientras que el 15% restante fueron 5 unidades de departamentos. Finalmente sólo 14 inmuebles presentaron 5 dormitorios, en todos los casos fueron casas.

En cuanto a las cocheras, se observaron tres valores extremos con una única observación en cada caso. Se trata de inmuebles con cocheras para 4, 6 y 8 vehículos. Por otra parte, debido a la gran cantidad de lotes presentes en el conjunto de datos, la mayoría de los inmuebles no tiene cochera. Del total de inmuebles que no tienen cochera, los lotes representan el 86,5% mientras que los departamentos representan el 10% (51 inmuebles) y las casas sólo el 3,5% (17 unidades). Si se analiza desde el punto de vista del tipo de inmueble, estos 17 inmuebles representan el 8,7% de las casas, mientras que los 51 departamentos corresponden al 51,5% de departamentos.

Siguiendo la línea de análisis a partir del tipo de inmueble, el 57,7% de las casas tiene 1 cochera (112 unidades), el 27,3% (53 unidades) tiene cochera para 2 vehículos y el 4,6% tiene para 3 autos (9 observaciones). Por otra parte, del total de inmuebles con cochera para 1 vehículo, 70,4% son casas, mientras que el resto son departamentos (47 unidades). En cuanto a los departamentos que sí tienen cocheras, el 47,5% (47 observaciones) tiene cochera para 1 vehículo, mientras que sólo el 1% (1 inmueble) tiene para 2 vehículos.

2.2 Análisis y transformaciones de la(s) variable(s) objetivo

Dado que el objetivo es modelar el precio de los inmuebles, es de esperarse que la distribución sea normal y asimétrica. En el gráfico 2.8 puede observarse el histograma y la función de kernel ajustada sobre los datos, ambos codificados en colores según el tipo de inmueble. Debido a esta forma característica, se optó por realizarle una transformación logarítmica. De esta manera, al momento de ajustar las regresiones, se puede interpretar la variación de las variables exógenas como la variación porcentual. Al mismo tiempo, permiten diferenciar mejor los rangos de valores y a que se asemeja una distribución normal simétrica. Esta distribución puede verse en el gráfico 2.9.

Alternativamente, la variable objetivo puede plantearse como el precio por metro cuadrado. De esta manera, se puede observar cómo varía el precio de un inmueble a medida que se agregan más metros cuadrados a un inmueble, al mismo tiempo que permite utilizar una unidad de medida más agnóstica. En efecto, una vez creada dicha variable -pm2-, podemos ver que tiene una distribución similar a la variable precio. Con lo cual también se optó por realizar la transformación logarítmica. Dichas distribuciones pueden observarse en los gráficos 2.10 y 2.11.

2.3 Análisis de correlaciones

Debido a la diversidad de intereses que se persiguieron en el estudio, se realizó una gran cantidad de transformaciones a las variables dependientes e independientes. Para poder realizar una apreciación si dichas transformaciones tenían sentido, se realizó una matriz de correlación entre las variables exógenas y las cuatro variables endógenas. En el gráfico 2.12 se puede observar el grado de correlación lineal entre las variables. Algunas variables se presentan duplicadas, en el sentido que se utilizaron la variables originales, por ejemplo, las discretas y por otra parte, sus dummies correspondientes. El objetivo de dicha transformación se explica en el Capítulo III, en la sección de Modelos modificados. Por su parte, las variables de latitud (*lat*) y longitud (*long*) muestran una correlación con las variables objetivo, aunque en este caso estamos capturando sólo una porción de dicha relación. En la matriz se establece una relación positiva con la latitud, es decir, mientras más al norte, más caras las propiedades y, al mismo tiempo, una correlación negativa con la longitud, en otras palabras, mientras más al oeste se ubique el inmueble, mayor será su precio. Dicho esto, es de esperarse que dicha relación sea no lineal, tal como se representará gráficamente con los mapas y polígonos en la sección de análisis geoespacial.

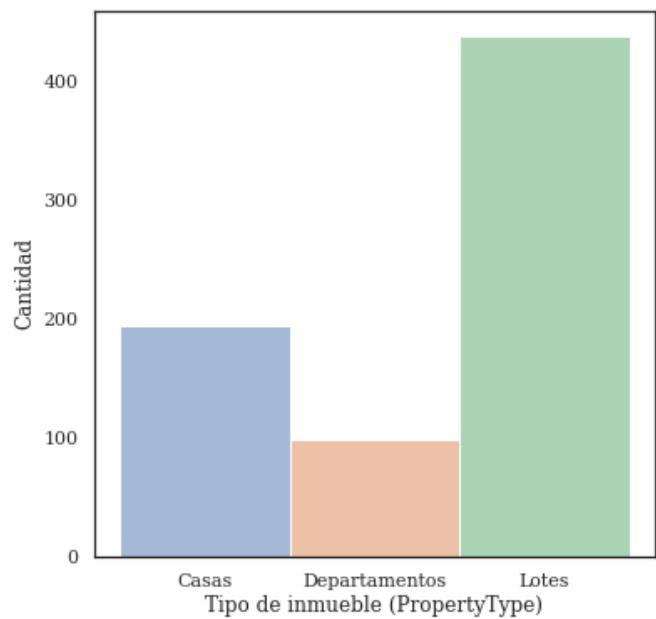
2.4 Ratios

Debido a que se observó una interacción entre las variables de metros construidos y metros totales, así como también con la cantidad de baños y de dormitorios, se optó por realizar dos variables que capturen dichas

relaciones: b_ratio y bed_bath_ratio , respectivamente. En el gráfico 2.13 se puede apreciar que a medida que aumenta la cantidad de dormitorios -eje x-, aumenta la cantidad de baños, los cuales están representados en colores. Al mismo tiempo, más del 40% de los inmuebles tienen 3 dormitorios.

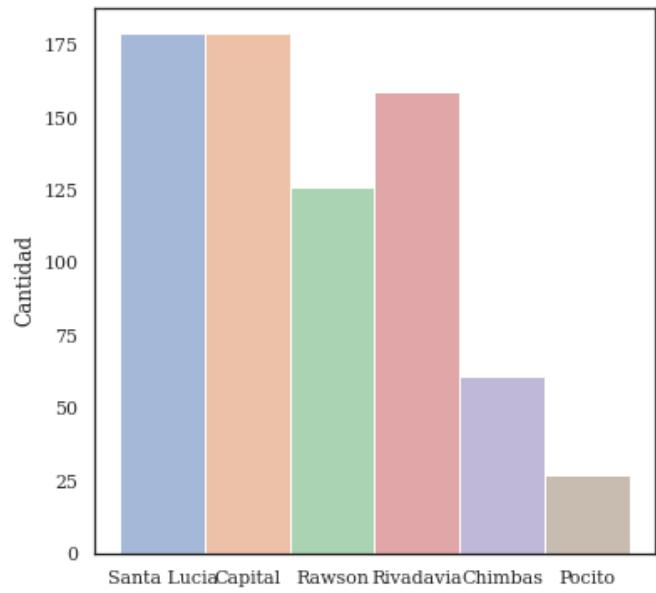
2.5 Dependencia espacial

A efectos de poder visualizar la distribución de precios en el espacio, se optó por realizar un gráfico de dispersión utilizando como base el mapa de la ciudad de San Juan y sus alrededores. De esta manera, el gráfico 2.14 representa los precios de los inmuebles en el espacio geográfico en escala logarítmica para facilitar la interpretación de las magnitudes de la misma. En el mismo se puede observar claramente el aumento de precios por metro cuadrado en las zonas céntricas, como es habitual en casi todas las ciudades. Asimismo, puede apreciarse rápidamente que en la zona al oeste de la capital, los precios son mayores que hacia el este. Para poder estimar los modelos que contemplan dependencia espacial, fue necesaria la construcción de la matriz de adyacencia utilizando la triangulación de Delaunay. En el gráfico 2.15 se pueden apreciar los polígonos de Voronoi (en celeste) y el grafo de adyacencia de las observaciones (en negro). En el grafo, cada nodo representa una observación y se conecta a través de aristas o bordes a sus puntos vecinos. Los polígonos, por su parte, son construidos partiendo las aristas en el punto medio entre dos nodos. Para visualizar este efecto, en el gráfico 2.16 se representa cada polígono (asociado a cada observación) con el color correspondiente a su precio por metro cuadrado, en escala logarítmica.



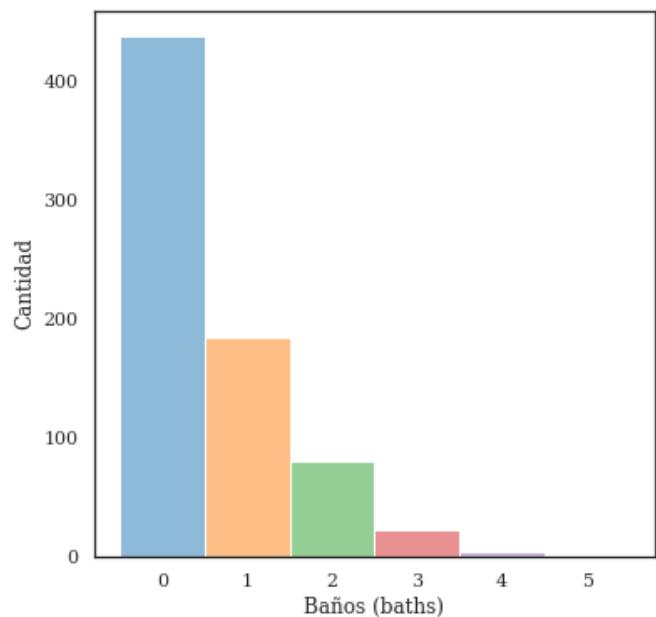
Fuente: Elaboración propia a partir de datos de compraensanjuan.com

Figura 2.1: Cantidad de avisos según tipo de inmueble.



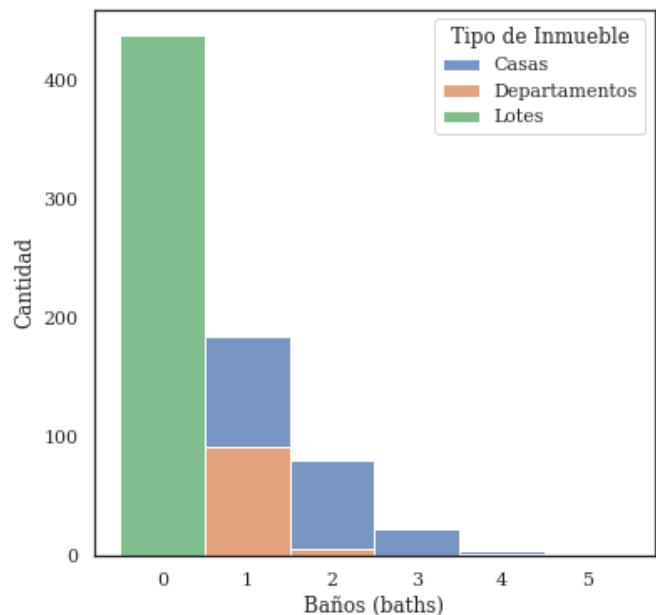
Fuente: Elaboración propia a partir de datos de compraensanjuan.com

Figura 2.2: Cantidad de avisos según localidad del inmueble.



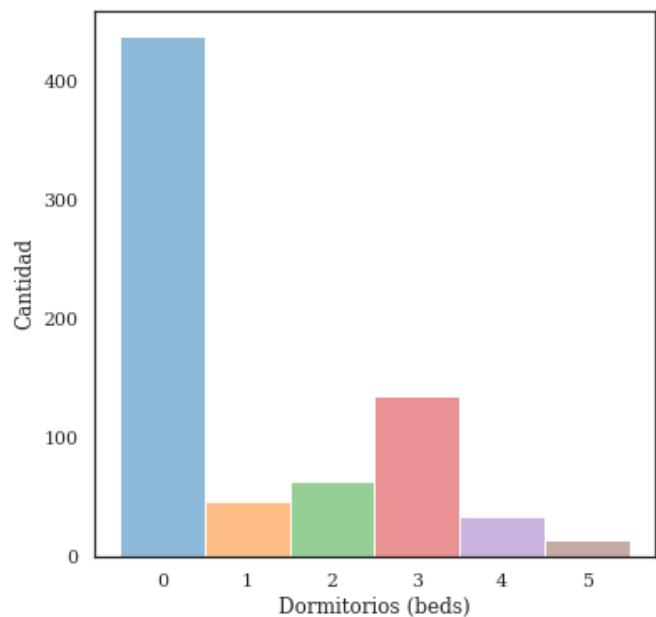
Fuente: Elaboración propia a partir de datos de compraensanjuan.com

Figura 2.3: Cantidad de avisos según número de baños.



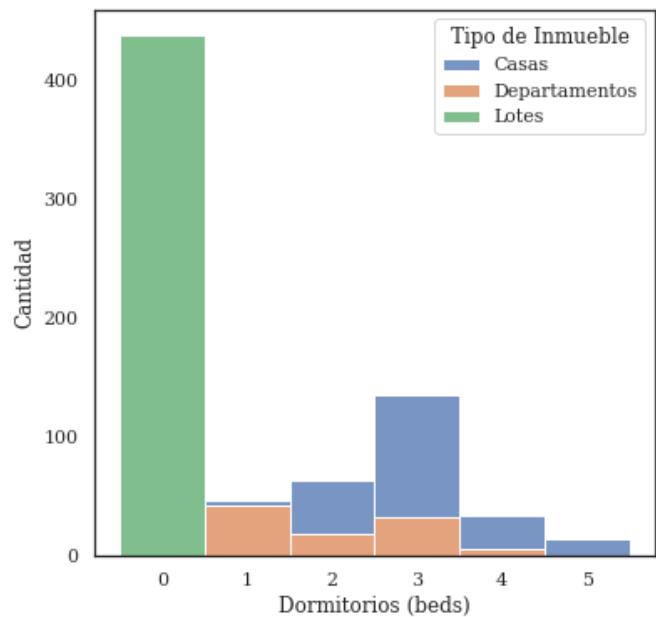
Fuente: Elaboración propia a partir de datos de compraensanjuan.com

Figura 2.4: Distribución de baños según tipo de inmueble.



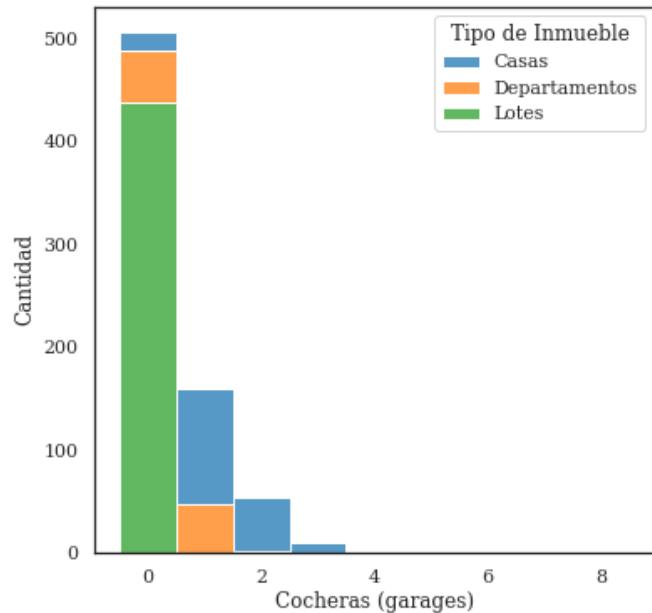
Fuente: Elaboración propia a partir de datos de compraensanjuan.com

Figura 2.5: Cantidad de avisos según número de Dormitorios.



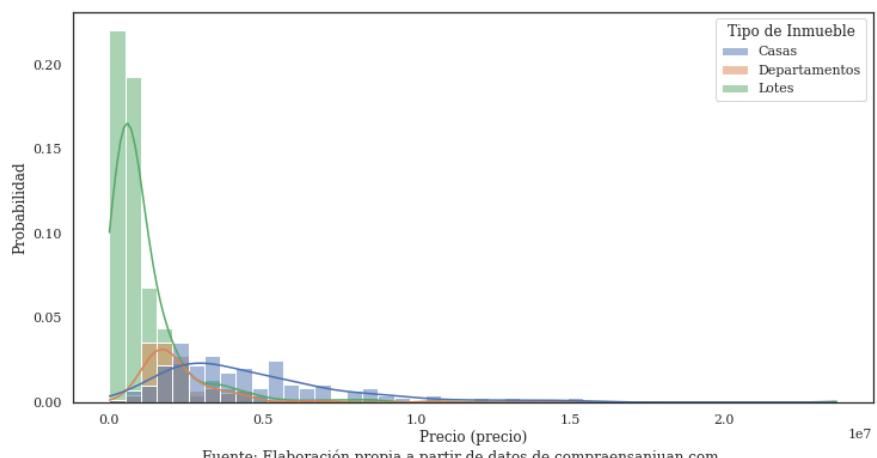
Fuente: Elaboración propia a partir de datos de compraensanjuan.com

Figura 2.6: Distribución de dormitorios según tipo de inmueble.



Fuente: Elaboración propia a partir de datos de compraensanjuan.com

Figura 2.7: Distribución de cocheras según tipo de inmueble.



Fuente: Elaboración propia a partir de datos de compraensanjuan.com

Figura 2.8: Distribución de precios de los inmuebles según tipo.

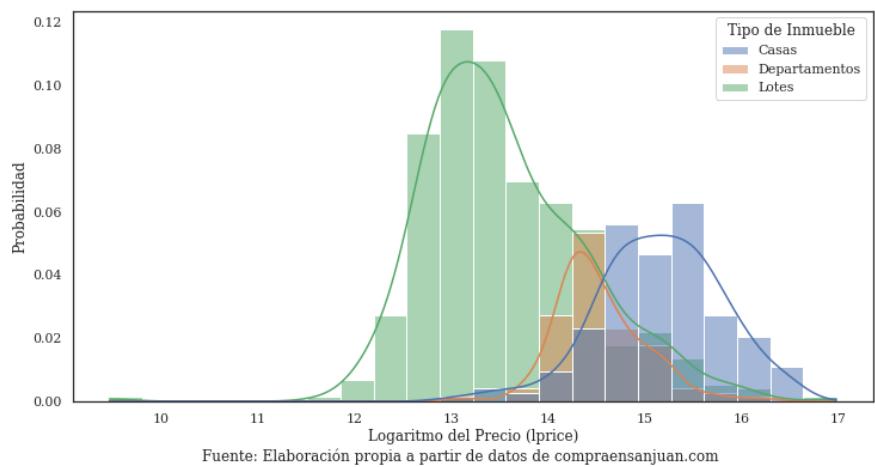


Figura 2.9: Distribución del logaritmo de precios de los inmuebles según tipo.

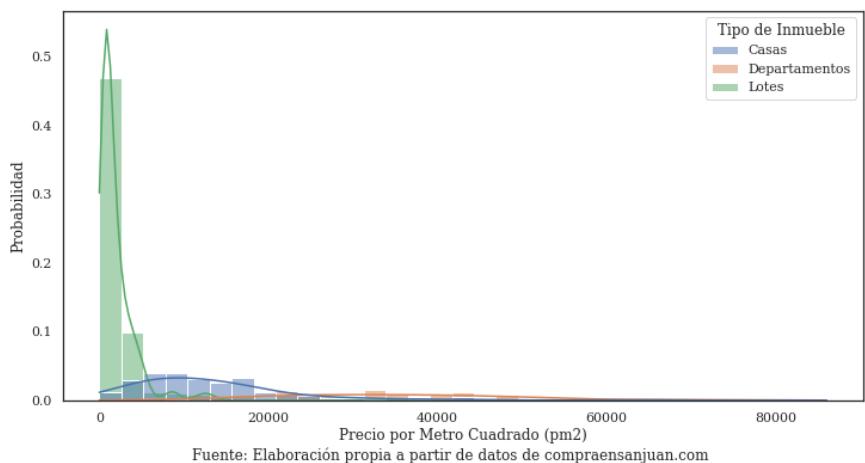


Figura 2.10: Distribución de valores de precio por metro cuadrado de inmuebles, según tipo.

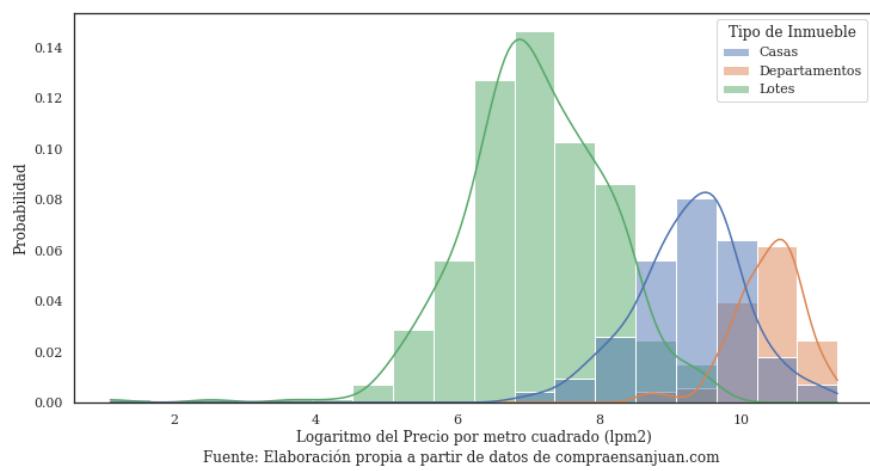
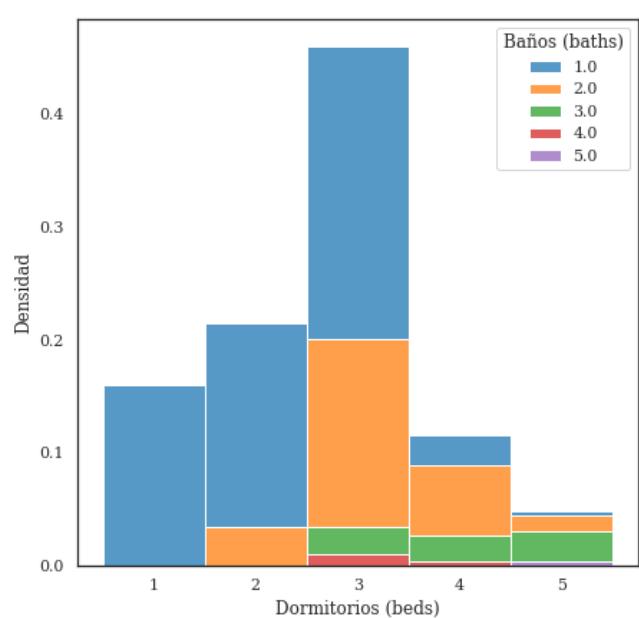


Figura 2.11: Distribución del logaritmo de precio por metro cuadrado de los inmuebles, según tipo.

	beds	baths	garages	bsqm
beds	0.59	0.68	0.49	0.68
baths	0.65	0.68	0.52	0.66
garages	0.56	0.56	0.31	0.48
bsqm	0.67	0.67	0.38	0.57
lat	0.069	0.041	0.078	0.16
long	-0.053	-0.11	-0.14	-0.27
Casas	0.56	0.61	0.18	0.44
artamentos	0.0063	0.16	0.71	0.57
Lotes	-0.51	-0.66	-0.66	-0.79
Capital	0.35	0.42	0.48	0.51
Chimbas	-0.16	-0.25	-0.16	-0.18
Pocito	-0.12	-0.15	-0.12	-0.22
Rawson	-0.15	-0.15	-0.17	-0.2
Rivadavia	0.034	0.1	0.038	0.16
Santa Lucia	-0.092	-0.15	-0.21	-0.27
car_0	-0.5	-0.59	-0.39	-0.58
car_1	0.25	0.38	0.32	0.46
car_2	0.4	0.36	0.16	0.26
car_3	0.092	0.12	0.056	0.09
car_4	0.12	0.073	-0.0095	0.014
car_6	0.093	0.067	-0.012	0.0095
car_8	0.19	0.087	0.026	0.042
bath_0	-0.51	-0.66	-0.66	-0.79
bath_1	0.11	0.31	0.55	0.59
bath_2	0.38	0.41	0.19	0.31
bath_3	0.39	0.28	0.1	0.16
bath_4	0.14	0.11	0.052	0.068
bath_5	0.14	0.079	0.11	0.065
bed_0	-0.51	-0.66	-0.66	-0.79
bed_1	-0.054	0.043	0.49	0.35
bed_2	0.11	0.21	0.2	0.26
bed_3	0.4	0.46	0.29	0.45
bed_4	0.22	0.23	0.13	0.2
bed_5	0.25	0.2	0.059	0.11
tsqm	0.0032	-0.023	-0.11	-0.34
tsqm2	-0.026	-0.047	-0.039	-0.22
bsqm2	0.56	0.47	0.18	0.32
b_ratio	0.35	0.49	0.86	0.81
_bath_ratio	0.47	0.57	0.68	0.72
	precio	lprice	pm2	lpm2

Fuente: Elaboración propia a partir de datos de compraensanjuan.com

Figura 2.12: Matriz de correlación de variables endógenas y exógenas.



Fuente: Elaboración propia a partir de datos de compraensanjuan.com

Figura 2.13: Proporción de baños según cantidad de dormitorios.

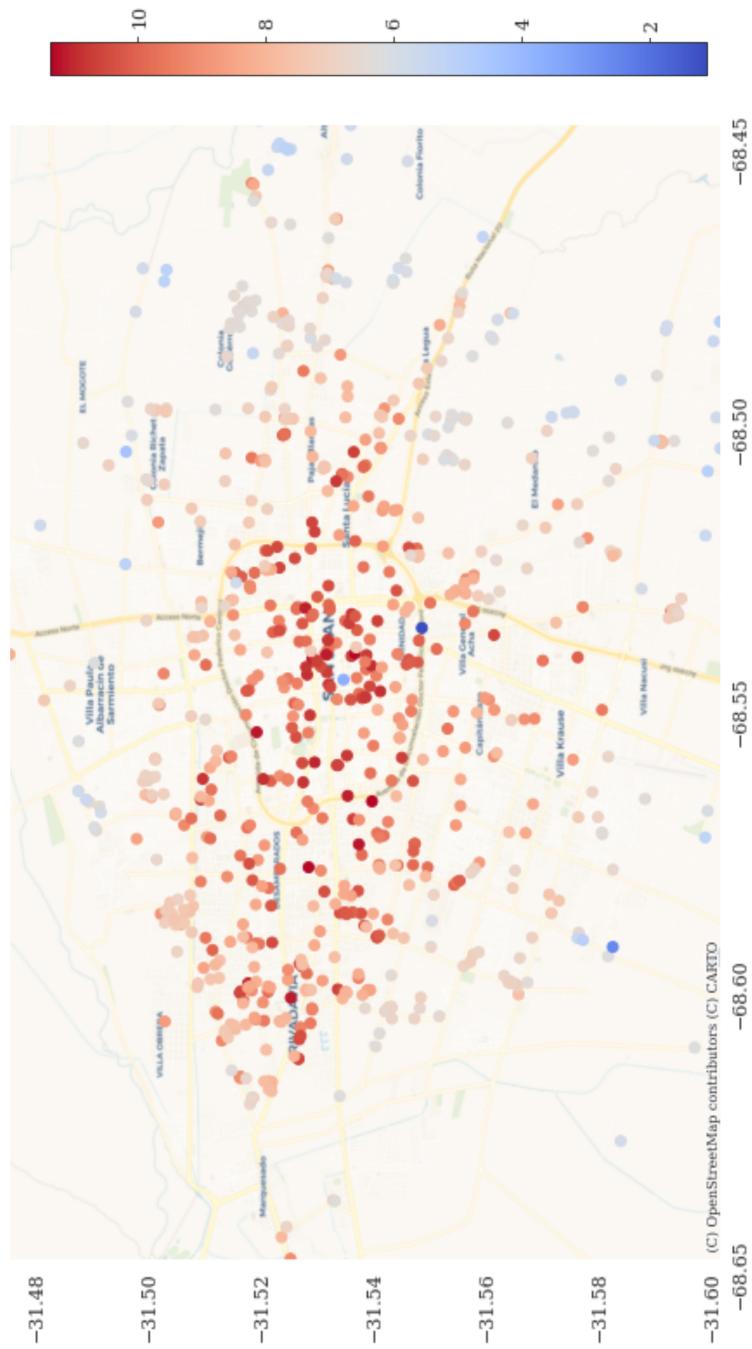


Figura 2.14: Mapa de dispersión coloreado en función del logaritmo de precio por metro cuadrado. Como puede observarse, no se trata de una distribución aleatoria, sino que se puede observar que hacia el centro del Gran San Juan, los precios aumentan.

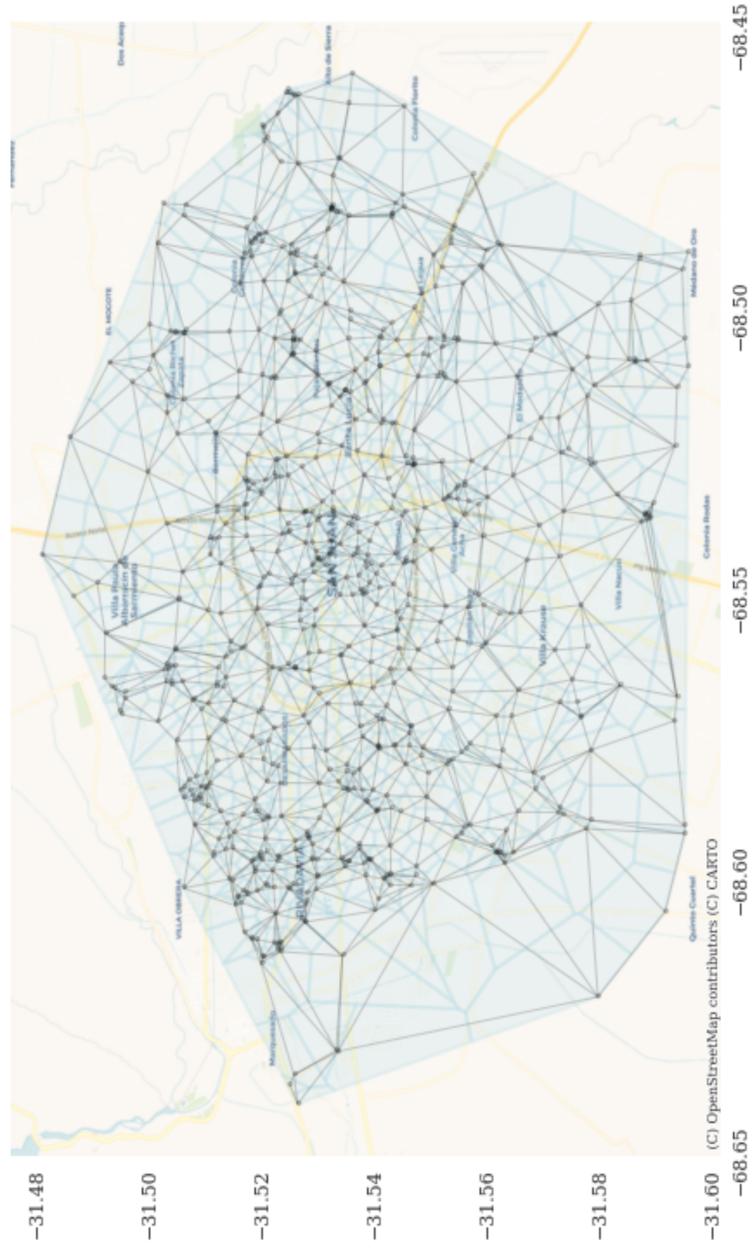


Figura 2.15: Polígonos de Voronoi. En la figura se pueden observar los puntos originales representados como nodos (vértices), y su relación de vecino con los polígonos adyacentes que se encuentran representados por bordes o aristas.

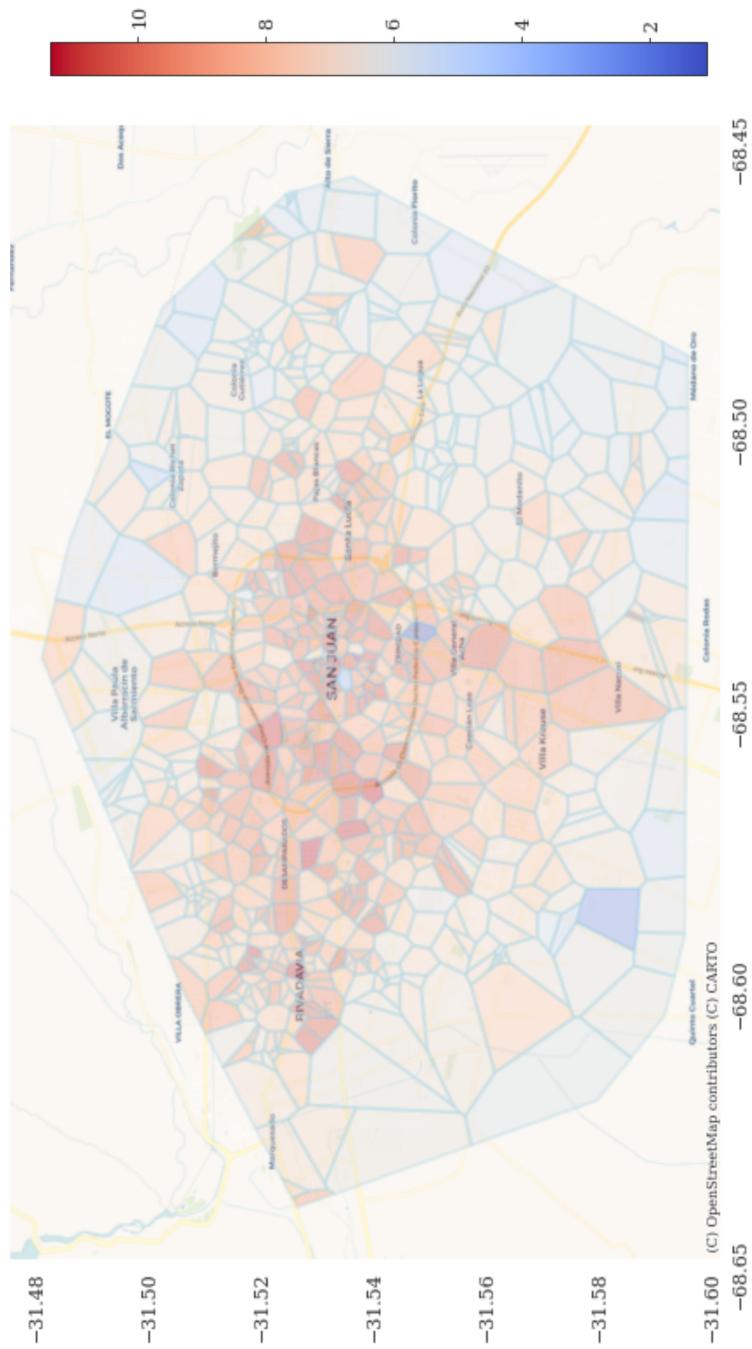


Figura 2.16: Mapa coroplético del logaritmo de precio por metro cuadrado. En esta representación, cada observación es extendida a un polígono de Voronoi para generalizar el área de influencia.

Capítulo 3

Especificación y estimación de modelos paramétricos

3.1 Especificación

3.1.1 MODELO BASE

Partiendo de la ecuación 1.2.2, procedemos a definir su forma funcional de la siguiente manera: $y = Z\beta + \varepsilon$ donde y es un vector de dimensiones $H \times 1$ de elementos $y - \ln p_h$ siendo p_h el precio de cada inmueble, Z es una matriz $H \times C$ de características del inmueble, β es un vector $C \times 1$ de los precios sombra de las características. En nuestro caso en particular, las características a considerar serán:

- Metros cuadrados de superficie total de la parcela, con el nombre *tsqm*.
- El cuadrado de Metros Cuadrados de superficie total, con el nombre

tsqm2.

- Metros cuadrados construidos, con el nombre *bsqm*.
- El cuadrado del total de metros cuadrados construidos *bsqm2*.
- Cantidad de Dormitorios *beds* , como variable continua discreta.
- Cantidad de Baños *baths*, como variable continua discreta.
- Cantidad de cocheras *garages* , como variable continua discreta.
- Tipo de propiedad, compuesta por las dummies: *Casas*, *Departamentos* y *Lotes*.
- Localidad del inmueble, compuesta por las dummies: *Chimbas*, *Pocito*, *Rawson*, *Rivadavia* y *SantaLuca*.

Con todas las variables construidas, se procedió a ajustar el modelo de regresión por Mínimos Cuadrados Ordinarios, en base a 731 observaciones disponibles en el dataset.

En base a los resultados obtenidos de la regresión (ver Tabla 3.1, podemos tomar nota de la significancia de las variables, así como también del ajuste general del modelo. Partiendo por éste el R^2 es de 0.559(0.55 ajustado), es decir, sólo explica dicho porcentaje de la variación de y . Respecto de las características incluidas en el modelo, se puede apreciar que hay dos de ellas que no son estadísticamente significativas al 95%: una de ellas es la cantidad de dormitorios *beds*, mientras que la otra es el cuadrado de la cantidad de metros cuadrados construidos *bsqm2*.

En ambos casos, esto podría sugerir que, en caso de que estas variables tengan información significativa -lo cual es bastante probable-, el tipo de relación no es lineal, por lo cual lo ideal sería tratar a *beds* como una variable categórica y crear las dummies correspondientes, mientras que la interpretación de *bsqm2*, podría sugerir que no es significativa respecto de la variable objetivo precio, pero sí quizás respecto de una transformación no lineal de la misma. En definitiva, en ambos casos nos encontraríamos con supuestos de no linealidad. Alternativamente, *bsqm2* podría verse influenciada por la magnitud de la unidad de medida de precio que se cuenta expresada en pesos a Junio de 2019.

3.1.2 MODELO BASE CON TRANSFORMACIÓN LOGARÍTMICA

Con la intención de capturar la variación porcentual en la variable dependiente, se puede realizar una transformación logarítmica sobre la misma. Realizar dicha transformación implica que la interpretación de los betas de la regresión pasan a estar medidos en puntos porcentuales de variación respecto de la variable objetivo, así como también introducimos no-linealidad

Cuadro 3.1: Regresión por MCO de 'precio', sin geolocalización y con variables continuas.

Model:	OLS	Adj. R-squared:	0.550	
Dependent Variable:	precio	AIC:	23062.4828	
Date:	2021-11-22 18:47	BIC:	23131.3990	
No. Observations:	731	Log-Likelihood:	-11516.	
Df Model:	14	F-statistic:	64.75	
Df Residuals:	716	Prob (F-statistic):	6.15e-117	
R-squared:	0.559	Scale:	2.8864e+12	
	Coef.	Std.Err.	t	P> t
Intercept	1001794.6909	489718.7083	2.0457	0.0412
Departamentos	-577266.1171	289089.8384	-1.9968	0.0462
Lotes	1011929.7035	480487.5601	2.1060	0.0355
Chimbas	-1913483.3892	267030.6478	-7.1658	0.0000
Pocito	-2085066.0239	369021.5508	-5.6503	0.0000
Rawson	-1540117.9699	210743.4251	-7.3080	0.0000
Rivadavia	-606423.6150	191922.3262	-3.1597	0.0016
Santa Lucia	-1084575.4259	194445.3649	-5.5778	0.0000
tsqm	256.1682	35.9391	7.1278	0.0000
tsqm2	-0.0028	0.0004	-6.4585	0.0000
bsqm	16586.3408	4905.4892	3.3812	0.0008
bsqm2	-14.4830	9.9110	-1.4613	0.1444
beds	-80205.9649	131241.1543	-0.6111	0.5413
baths	1018250.0355	199791.8962	5.0966	0.0000
garages	552066.7752	131700.1101	4.1918	0.0000
Omnibus:	648.787	Durbin-Watson:	1.878	
Prob(Omnibus):	0.000	Jarque-Bera (JB):	33919.168	
Skew:	3.706	Prob(JB):	0.000	
Kurtosis:	35.538	Condition No.:	4398056683	

Fuente: Elaboración propia a partir de datos de compraensanjuan.com, 2019.

en el modelo. En efecto, en la Tabla 3.2 se pueden observar los nuevos valores de los betas, así como también el hecho de que ahora $bsqm2$ es estadísticamente significativa con un $\alpha = 5\%$, tal como se postuló en la sección anterior. Adicionalmente puede observarse que el R^2 ajustado aumentó a 0.632. Respecto de la interpretación de los coeficientes, por ejemplo, el beta asociado a la localidad de Rivadavia Rivadavia, en la regresión con la variable precio toma un valor de -60,6423.615, es decir en promedio 60,6423.615 menos que un inmueble de Capital. Mientras que en el segundo caso, toma un valor de -0.2021, es decir un 20% menos en promedio. Nuevamente se puede observar que hay gran disparidad entre los coeficientes, lo cual tiene que ver con las magnitudes de medida de cada una de las variables. Esto también explica el valor extremo que toma el Número de Condición (4398056683).

3.2 Modelos modificados

3.2.1 PRECIO POR METRO CUADRADO (PM2) COMO VARIABLE DEPENDIENTE

Una alternativa bastante interesante a analizar, consiste en tratar de estimar el precio del metro cuadrado $pm2$ del inmueble, en vez del precio total del mismo, es decir: $pm2 = \frac{precio}{tsqm}$ {#eq:pm2}

En la Tabla 3.3 se pueden observar los resultados de utilizar $pm2$ como variable dependiente. En la misma se destaca que variables que previamente eran significativas al 95%, ahora no lo son, tal es el caso de $tsqm$, $tsqm2$ y $garages$, mientras que otras variables exógenas como $bsqm2$ o $beds$ ahora sí lo son, al 10% y al 5%, respectivamente. El hecho de que $tsqm$ y $tsqm2$ no sean estadísticamente significativas en este esquema tiene sentido ya que lo más probable es que no se trate de una relación lineal o cuadrática, ya que la misma variable $tsqm$ se utilizó para transformar el precio y construir $pm2$. Con este resultado se puede concluir que no se está filtrando información al modelo.

En cuanto al poder descriptivo del modelo, podemos ver una mejora considerable en tanto en el valor de R^2 como en $R^2_{ajustado}$, que aumentan de 0.559 y 0.55a 0.708 y 0.702, respectivamente; es decir, se observa una mejora de 26, 5p.p. y de 27, 6p.p. en cada caso.

Cuadro 3.2: Regresión por MCO de 'lprice', sin geolocalización y con variables continuas.

Model:	OLS	Adj. R-squared:	0.632	
Dependent Variable:	lprice	AIC:	1404.5282	
Date:	2021-11-22 17:57	BIC:	1473.4444	
No. Observations:	731	Log-Likelihood:	-687.26	
Df Model:	14	F-statistic:	90.70	
Df Residuals:	716	Prob (F-statistic):	5.41e-148	
R-squared:	0.639	Scale:	0.39188	
	Coef.	Std.Err.	t	P > t
Intercept	14.3944	0.1804	79.7719	0.0000
Departamentos	-0.2928	0.1065	-2.7486	0.0061
Lotes	-0.4387	0.1770	-2.4782	0.0134
Chimbas	-1.1291	0.0984	-11.4755	0.0000
Pocito	-1.0620	0.1360	-7.8104	0.0000
Rawson	-0.6674	0.0777	-8.5948	0.0000
Rivadavia	-0.2021	0.0707	-2.8583	0.0044
Santa Lucia	-0.5513	0.0716	-7.6954	0.0000
tsqm	0.0001	0.0000	9.3906	0.0000
tsqm2	-0.0000	0.0000	-8.7855	0.0000
bsqm	0.0050	0.0018	2.7501	0.0061
bsqm2	-0.0000	0.0000	-2.1373	0.0329
beds	0.0130	0.0484	0.2686	0.7883
baths	0.2031	0.0736	2.7588	0.0059
garages	0.1065	0.0485	2.1946	0.0285
Omnibus:	82.214	Durbin-Watson:	1.701	
Prob(Omnibus):	0.000	Jarque-Bera (JB):	627.788	
Skew:	-0.075	Prob(JB):	0.000	
Kurtosis:	7.537	Condition No.:	4398056683	

Fuente: Elaboración propia a partir de datos de compraensanjuan.com, 2019.

3.2.2 LOGARITMO DEL PRECIO POR METRO CUADRADO

LPM2 COMO VARIABLE DEPENDIENTE.

En línea con la sección anterior, en ésta se busca modelar el logaritmo del precio por metro cuadrado -*lpm2*- a partir de las mismas variables exógenas que en los casos anteriores. Como se propuso previamente, el hecho que las variables *tsqm* y su cuadrado no sean estadísticamente significativas, da lugar a sospechar que puede deberse a una relación de no linealidad. Tal como se observa en la 3.4, las variables *tsqm* y *tsqm2* pasan a ser estadísticamente significativas aunque con un coeficiente muy cercano a cero, debido a la unidad de medida de la misma, al punto que *tsqm2* es redondeada a cero por la librería *statsmodels*. Si bien, se observa que la cantidad de dormitorios -*beds*- deja de ser estadísticamente significativa la razón es relativamente sencilla: aparentemente tanto la cantidad de metros -*tsqm*-, como su transformación cuadrática -*tsqm2*-, contienen la misma información que la variable dormitorios -*beds*.

Cuadro 3.3: Regresión por MCO de 'pm2', sin geolocalización y con variables continuas.

Model:	OLS	Adj. R-squared:	0.702
Dependent Variable:	pm2	AIC:	15213.7604
Date:	2021-11-22 19:13	BIC:	15282.6766
No. Observations:	731	Log-Likelihood:	-7591.9
Df Model:	14	F-statistic:	123.8
Df Residuals:	716	Prob (F-statistic):	2.62e-180
R-squared:	0.708	Scale:	6.2713e+07

	Coef.	Std.Err.	t	P> t
Intercept	21682.9033	2282.6749	9.4989	0.0000
Departamentos	18548.9113	1347.5044	13.7654	0.0000
Lotes	-12486.8167	2239.6467	-5.5754	0.0000
Chimbas	-9554.3840	1244.6822	-7.6762	0.0000
Pocito	-8616.4696	1720.0818	-5.0093	0.0000
Rawson	-8563.9327	982.3164	-8.7181	0.0000
Rivadavia	-6225.0700	894.5876	-6.9586	0.0000
Santa Lucia	-8207.5256	906.3480	-9.0556	0.0000
tsqm	-0.1454	0.1675	-0.8677	0.3858
tsqm2	0.0000	0.0000	0.7724	0.4401
bsqm	-64.5413	22.8654	-2.8227	0.0049
bsqm2	0.0790	0.0462	1.7108	0.0876
garages	407.6286	613.8800	0.6640	0.5069
baths	5738.3234	931.2692	6.1618	0.0000
beds	-1870.8038	611.7407	-3.0582	0.0023

Omnibus:	461.609	Durbin-Watson:	1.889
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7929.654
Skew:	2.531	Prob(JB):	0.000
Kurtosis:	18.321	Condition No.:	4398056683

Fuente: Elaboración propia a partir de datos de compraensanjuan.com, 2019.

3.2.3 ESTABLECIENDO RELACIÓN NO LINEAL CON VARIABLES DISCRETAS Y UTILIZANDO $lpm2$ COMO VARIABLE DEPENDIENTE.

En base a los resultados obtenidos y las mejoras observadas, en esta sección se procederá a combinar todas las transformaciones previamente realizadas, así como también una relación no lineal con las variables discretas como la cantidad de baños (baths), dormitorios (beds) y cocheras (garages). De esta manera, se generarán $n - 1$ variables *dummy* por cada variable discreta.

Debido a la gran cantidad de variables dicotómicas, cabe destacar que el intercepto contiene los efectos de las dummies bed0, bath0, garages0, Capital y Casas, que fueron eliminadas para evitar multicolinealidad. Los resultados de dicho modelo se encuentran en la Tabla 3.5, donde se puede observar claramente que el R^2 aumentó a 0.843 (0.837). Debido a que la mayoría de las variables exógenas son estadísticamente significativas, se detalla a continuación las que no lo son:

- En el caso de las variables *garages1*, *garages3*, *garages4*, *garages6* y *garages8* no se puede rechazar la hipótesis nula de que sean estadísticamente distintos de cero, a diferencia de *garages2* que sí lo es, con un coeficiente de 0.3974(0.1397s.e.) y un valor $t = 2.8457$. Esto podría indicar la posibilidad de una interacción entre garages y el tipo de inmueble *PropertyType* y/o la localidad \$\$.
- *bsqm2* tampoco es significativa, lo que sugiere que la relación entre *bsqm* y el logaritmo del precio por metro cuadrado es simplemente

Cuadro 3.4: Regresión por MCO de 'lpm2', sin geolocalización y con variables continuas.

Model:	OLS	Adj. R-squared:	0.819	
Dependent Variable:	lpm2	AIC:	1503.7073	
No. Observations:	731	BIC:	1572.6235	
Df Model:	14	Log-Likelihood:	-736.85	
Df Residuals:	716	F-statistic:	237.3	
R-squared:	0.823	Prob (F-statistic):	1.06e-257	
	Coef.	Std.Err.	t	P> t
Intercept	9.6328	0.1931	49.8825	0.0000
PropertyTypeDepartamentos]	0.8936	0.1140	7.8388	0.0000
PropertyTypeLotes]	-1.5692	0.1895	-8.2823	0.0000
Chimbas]	-1.1213	0.1053	-10.6489	0.0000
Pocito]	-1.4413	0.1455	-9.9045	0.0000
Rawson]	-0.9767	0.0831	-11.7530	0.0000
Rivadavia]	-0.2931	0.0757	-3.8735	0.0001
Santa Lucia]	-1.0327	0.0767	-13.4679	0.0000
tsqm	-0.0001	0.0000	-10.1971	0.0000
tsqm2	0.0000	0.0000	5.3335	0.0000
bsqm	-0.0019	0.0019	-0.9649	0.3349
bsqm2	0.0000	0.0000	0.4166	0.6771
garages	0.0559	0.0519	1.0771	0.2818
baths	0.2353	0.0788	2.9864	0.0029
beds	-0.0203	0.0518	-0.3930	0.6945
Omnibus:	0.453	Durbin-Watson:	1.539	
Prob(Omnibus):	0.797	Jarque-Bera (JB):	0.318	
Skew:	0.021	Prob(JB):	0.853	
Kurtosis:	3.093	Condition No.:	4398056683	

Fuente: Elaboración propia a partir de datos de compraensanjuan.com, 2019.

lineal.

- *Departamentos* no resultó ser estadísticamente significativa y esto puede deberse a que comparta información (es decir, que tenga correlación) con otras variables, como pueden ser *b_ratio*, *bed1* y *bath1*.

3.3 Modelo Spatial Lag o SAR (Simultaneous Autoregressive-Regressive)

Como se anticipó previamente, el modelo de Spatial Lag, considera que la información de las observaciones vecinas aporta información a la hora de construir el precio del inmueble en cuestión. Esto significa que el precio del inmueble i se construye, en parte, gracias a las características del inmueble j –es decir WX_j- , gracias al precio de dicho inmueble –es decir WY_j- , o bien en los residuos del modelo –es decir W_u- , donde W es la matriz de pesos espaciales. Cabe destacar que dicha matriz es calculada en base a la distancia entre una observación y sus vecinos, normalizada a 1, ya sea dividiendo las filas para que tomen dicho valor o las columnas. Nótese que el motivo por el cual se incorpora la matriz de pesos espaciales al cálculo, tiene que ver con la limitación que se encuentra a la hora de estimar las interacciones de los parámetros de las observaciones: no es posible calcular todas las interacciones para cada uno de los parámetros ya que existen n^2 parámetros para n observaciones. De esta manera, el problema se restringe a un único parámetro de dependencia espacial (λ) y se calcula sólo a partir de las observaciones vecinas.

3.3.1 MATRIZ DE PESOS ESPACIALES

De acuerdo a la especificación modelo, la forma de incorporación del concepto de vecindad de dos observaciones es representada por la matriz de pesos espaciales. Para poder estimarla, se procedió a la creación de polígonos de Voronoi y posteriormente se calculó la contigüidad a partir del criterio de

Cuadro 3.5: Regresión por MCO de 'lpm2', sin geolocalización y con variables dummies.

Model:	OLS	Adj. R-squared:	0.837	
Dependent Variable:	lpm2	AIC:	1437.6488	
Date:	2021-11-22 19:37	BIC:	1561.6980	
No. Observations:	731	Log-Likelihood:	-691.82	
Df Model:	26	F-statistic:	145.7	
Df Residuals:	704	Prob (F-statistic):	6.38e-263	
R-squared:	0.843	Scale:	0.40356	
	Coef.	Std.Err.	t	P> t
Intercept	7.1121	0.1727	41.1929	0.0000
garages1	0.0899	0.1060	0.8482	0.3966
garages2	0.3974	0.1397	2.8457	0.0046
garages3	-0.0320	0.2395	-0.1338	0.8936
garages4	1.0171	0.6911	1.4717	0.1416
garages6	0.5298	0.6773	0.7822	0.4344
garages8	0.9300	0.7789	1.1940	0.2329
baths1	0.8429	0.1458	5.7795	0.0000
baths2	1.1393	0.1422	8.0121	0.0000
baths3	1.4565	0.1803	8.0793	0.0000
baths4	1.3351	0.3223	4.1423	0.0000
baths5	1.4589	0.5930	2.4603	0.0141
beds1	1.3075	0.1262	10.3562	0.0000
beds2	1.4396	0.1095	13.1425	0.0000
beds3	1.2795	0.1039	12.3140	0.0000
beds4	1.2425	0.1360	9.1388	0.0000
beds5	0.9637	0.1988	4.8476	0.0000
Departamentos	-0.0625	0.1682	-0.3719	0.7101
Lotes	0.8794	0.1698	5.1782	0.0000
Chimbas	-1.0453	0.1015	-10.3018	0.0000
Pocito	-1.3657	0.1394	-9.8001	0.0000
Rawson	-0.9187	0.0809	-11.3544	0.0000
Rivadavia	-0.2452	0.0748	-3.2763	0.0011
Santa Lucia	-0.9298	0.0755	-12.3075	0.0000
tsqm	-0.0001	0.0000	-10.3293	0.0000
tsqm2	0.0000	0.0000	5.2424	0.0000
bsqm	-0.0044	0.0020	-2.1585	0.0312
bsqm2	0.0000	0.0000	0.5386	0.5904
b_ratio	1.6997	0.1983	8.5730	0.0000
Omnibus:	1.272	Durbin-Watson:	1.485	
Prob(Omnibus):	0.529	Jarque-Bera (JB):	1.142	
Skew:	-0.024	Prob(JB):	0.565	
Kurtosis:	3.188	Condition No.:	10024944593774462	

Fuente: Elaboración propia a partir de datos de compraensanjuan.com, 2019.

Reina (Queen), ya que sólo se consideró relevante que sean o no vecinas las distintas observaciones. Una vez estimada y normalizada la matriz de pesos espaciales, se procedió al ajuste del modelo SAR por el método de máxima verosimilitud.

3.3.2 SAR POR MÉTODO DE MÁXIMA VERO SIMILITUD

De acuerdo al ajuste del modelo realizado con esta metodología, se puede observar en la Tabla 3.6 que la cantidad de dormitorios presenta una interacción no lineal estadísticamente significativa, ya que la variable beds_sq obtuvo un coeficiente de $-0.3653(s.e.0.0526)$, mientras que la variable beds obtuvo un beta de $0.5669(s.e.0.1786)$. Con respecto a la cantidad de cocheras, si bien son estadísticamente significativas, el tipo de relación que se observa es lineal, ya que sólo fue significativa garages con un coeficiente de $0.1711(s.e.0.0526)$, a diferencia de $garages_2$ que no presentó un coeficiente significativo.

En cuanto a las variables de mensura, se puede observar una relación cuadrática con respecto a la cantidad de metros cuadrados totales al cuadrado, $tsqm_2$, con coeficiente $0.2499(s.e.0.0498)$ y coeficiente $-0.6147(s.e.0.0498)$ para la cantidad de metros cuadrados totales, es decir, $tsqm$. Con respecto a la cantidad de metros cuadrados construidos, se observa una relación lineal con un coeficiente significativo para $bsqm$ de valor $-0.2896(s.e.0.1129)$, pero con coeficiente cuadrático no significativo. También se observó un coeficiente estadísticamente significativo para la variable de proporción de construcción b_ratio con un β de $0.5501(s.e.0.0401)$.

Finalmente, de acuerdo a las restricciones impuestas, podemos decir que existe autocorrelación espacial positiva respecto de los precios de las otras observaciones. Dicho efecto está capturado por W_lpm2 y presenta un coeficiente de 0.4868(*s.e.*0.0193). Este modelo nos permite estimar un *pseudo-R²* de 0.8923 y un *pseudo-R²espacial* de 0.8465.

Cuadro 3.6: Modelo SAR con matriz de pesos espaciales de contiguidad de Queen.

Model:	OLS	Presudo R-squared:	0.8923	
Dependent Variable:	lpm2	Spatial Pseudo R-squared:	0.8465	
Mean dependent var	8.1251	AIC:	1173.428	
S.D. dependent var	1.5757	Log-Likelihood:	-572.714	
Df Model:	14	Schwarz criterion:	1237.750	
Df Residuals:	717	S.E of regression:	0.517	
Sigma-square ML	0.267			
	Coef.	Std.Err.	z	P> t
Intercept	4.1704	0.1581	26.3753	0.0000
beds	0.5669	0.1786	3.1739	0.0015
baths	0.2896	0.2196	1.3183	0.1873
garages	0.1711	0.0526	3.2479	0.0011
tsqm	-0.2868	0.1129	-2.5409	0.0110
bsqm	-0.6147	0.0502	-12.2406	0.0000
tsqm2	0.2499	0.0498	5.0085	0.0000
bsqm2	0.0328	0.0715	0.4585	0.6465
b_ratio	0.5501	0.0401	13.7072	0.0000
bed_bath_ratio	0.1134	0.0780	1.4535	0.1460
beds_sq	-0.3653	0.1100	-3.3196	0.0009
baths_sq	-0.1214	0.1074	-1.1307	0.2581
garages_sq	-0.0385	0.0403	-0.9566	0.3387
W_lpm2	0.4868	0.0193	25.1787	0.0000

Fuente: Elaboración propia a partir de datos de compraensanjuan.com, 2019.

3.3.3 MODELO SLX

Esta versión es considerada la más simple y propone que el precio de un inmueble depende no sólo de sus características, sino también de las características de los inmuebles vecinos. La forma matemática es bastante sencilla y se define como: $\log pm2_i = \beta_0 + \beta \times X_i + \rho W \times X_j + \varepsilon$ {#eq:slx}

El vector de parámetros ρ es el que tendremos que evaluar si es significativo o no, con las mismas pruebas de significancia establecidos para la estimación de los betas. De esta manera, se evita establecer un modelo

endógeno como el observado en el modelo SAR, lo que nos da flexibilidad a la hora de realizar predicciones, tiende a reducir la capacidad descriptiva de la varianza observada en la variable objetivo.

Como se puede observar en la Tabla 3.7, incluir información de las características de los inmuebles vecinos no aporta mucho a la descripción de la variación de la variable objetivo. El R^2 pasó de 0.843(0.837) a 0.875(0.865) incluyendo la dimensión espacial. No sólo eso, sino que se puede apreciar la penalización sobre el $R^2 - Ajustado$, ya que si bien el modelo había explicado 0.5p.p. más, en base a la cantidad de variables no significativas incluidas, el R^2 Ajustado aumentó sólo 0.28p.p.

Entre las nuevas variables incluidas, podemos observar que las únicas estadísticamente significativas para un α de 5% son: *tsqm*, *tsqm2*, *bsqm*, *b_ratio*, *Chimbas*, *Pocito*, *Rawson*, *Rivadavia*, *car_2*, *bath_3*, *Lotes*, *lag_car_3*, *lag_car_4*, *lag_bed_5*, *lag_bath_3*, *lag_bath_5*, *lag_Lotes*, *lag_tsqm*, *lag_Rawson*, *lag_SantaLucia* y *lag_tsqm2*. El resto de las variables no son estadísticamente significativas.

Cuadro 3.7: Modelo SAR con matriz de pesos espaciales de contiguidad de Queen.

Model:	OLS	Adj. R-squared:	0.865	
Dependent Variable:	y	AIC:	1328.8715	
No. Observations:	731	BIC:	1581.5642	
Df Model:	54	Log-Likelihood:	-609.44	
Df Residuals:	676	F-statistic:	87.53	
R-squared:	0.875	Prob (F-statistic):	9.47e-268	
	Coef.	Std.Err.	t	P> t
Intercept	14.3454	2.0783	6.9025	0.0000
bsqm	-0.3289	0.1404	-2.3431	0.0194
Departamentos	-0.0326	0.1606	-0.2030	0.8392
Lotes	-2.3713	1.0400	-2.2801	0.0229
Chimbas	-0.6160	0.1993	-3.0917	0.0021
Pocito	-0.7010	0.2444	-2.8675	0.0043
Rawson	-0.2926	0.1474	-1.9849	0.0476
Rivadavia	-0.3222	0.1465	-2.1998	0.0282
Santa Lucia	-0.1418	0.1511	-0.9385	0.3483
car_2	0.2347	0.1015	2.3113	0.0211
car_3	0.0185	0.2089	0.0884	0.9296
car_4	0.8340	0.6267	1.3308	0.1837
car_6	0.6994	0.6184	1.1310	0.2584
car_0	-0.1656	0.0999	-1.6576	0.0979
car_8	0.5178	0.7226	0.7165	0.4739
bath_2	0.6697	0.3760	1.7810	0.0754
bath_3	1.4979	0.6535	2.2920	0.0222
bath_4	1.7573	1.0432	1.6846	0.0925
bath_5	1.8335	1.1847	1.5476	0.1222
bed_2	-0.4402	0.5099	-0.8633	0.3883
bed_3	-0.8439	0.7074	-1.1930	0.2333
bed_4	-1.1551	0.8804	-1.3119	0.1900
bed_5	-1.5304	1.0217	-1.4980	0.1346
tsqm	-0.6292	0.0578	-10.8875	0.0000
tsqm2	0.2714	0.0571	4.7558	0.0000
bsqm2	0.0340	0.0902	0.3764	0.7067
b_ratio	0.5557	0.0685	8.1153	0.0000
bed_bath_ratio	-0.4280	0.3466	-1.2348	0.2173
lag_bsqm	0.2011	0.3718	0.5407	0.5889
lag_Departamentos	0.4222	0.3460	1.2204	0.2228
lag_Lotes	-5.3339	2.6435	-2.0178	0.0440

lag_Chimbas	-0.3095	0.2416	-1.2813	0.2005
lag_Rawson	-0.5984	0.1793	-3.3369	0.0009
lag_Rivadavia	0.1296	0.1725	0.7515	0.4526
lag_Santa Lucia	-0.7262	0.1803	-4.0283	0.0001
lag_car_2	-0.0332	0.2459	-0.1349	0.8927
lag_car_3	-1.1801	0.4796	-2.4604	0.0141
lag_car_4	3.9414	1.8787	2.0979	0.0363
lag_car_6	-0.9383	1.4885	-0.6304	0.5287
lag_car_0	-0.1251	0.2492	-0.5021	0.6158
lag_car_8	-1.4180	2.2695	-0.6248	0.5323
lag_bath_2	1.7141	0.9111	1.8814	0.0603
lag_bath_3	3.4931	1.6432	2.1258	0.0339
lag_bath_4	5.0116	2.6347	1.9022	0.0576
lag_bath_5	6.2346	3.0167	2.0667	0.0391
lag_bed_2	-2.2368	1.2754	-1.7538	0.0799
lag_bed_3	-3.3610	1.7693	-1.8997	0.0579
lag_bed_4	-4.0964	2.2018	-1.8605	0.0633
lag_bed_5	-6.2597	2.5554	-2.4496	0.0146
lag_tsqm	-0.4907	0.1192	-4.1183	0.0000
lag_tsqm2	0.4642	0.1236	3.7557	0.0002
lag_bsqm2	0.0028	0.2560	0.0109	0.9913
lag_b_ratio	-0.1458	0.1505	-0.9689	0.3330
lag_bed_bath_ratio	-1.6478	0.8656	-1.9036	0.0574

Fuente: Elaboración propia a partir de datos de compraensanjuan.com, 2019.

Capítulo 4

Especificación y estimación de modelos no paramétricos

4.1 Introducción

Tal como se describió previamente, en esta sección se procederá a la implementación de los algoritmos de *Random Forest* y *Gradient Boosting*, la optimización de sus parámetros utilizando validación cruzada y su poder de generalización con un set de testeo. En el caso del primer algoritmo, se procederá al ajuste del modelo utilizando la librería *Scikit-Learn*¹ de python debido a su simplicidad de uso, así como también a su popularidad. En cuanto al modelo de Gradient Boosting, se procederá a su ajuste a través de la librería *CatBoost*² ya que la misma no requiere de optimización de hiper-parámetros y es una de las versiones más eficientes en términos computacionales para el ajuste del modelo. Tanto el algoritmo de RandomForest

¹Pedregosa et al. (2011)

²Prokhorenkova, Liudmila et al. (2018)

como CatBoost nos permiten crear subdivisiones en el espacio de las características, generando así, interacciones no lineales entre las mismas. Esto nos permite arrojar por la borda toda formalidad de especificación del problema, ya que no se plantea la existencia de una forma funcional que exprese la relación de las variables, sino que simplemente se intenta predecir valores de la variable objetivo en base a los datos observados. Es por esta razón que es extremadamente importante dividir el dataset de las observaciones en al menos dos grupos: el conjunto de entrenamiento y el conjunto de validación. De esta forma podemos comprobar que el modelo generaliza bien a datos no observados.

4.2 Métodos de conjuntos secuenciales

4.2.1 CATBOOST SIN VARIABLES AUTOCORRELACIONADAS

Para obtener los mejores resultados se realizó una búsqueda de hiper-parámetros donde el espacio de posibles valores estuvo dado por: - Máxima profundidad de los árboles : 2, 4, 10 ó sin límite. - Tasa de aprendizaje: 0.001, 0.01, 0.1 ó 1. - Cantidad de estimadores: 50, 200, 500 ó 1000.

Luego de ajustar los tres pliegues de las 64 combinaciones posibles, la mejor combinación de hiper-parámetros resultó ser: máxima profundidad de 4 hojas, tasa de aprendizaje de 0.1 y 500 estimadores. Con esta configuración se obtuvo un R^2 de 0.977 en entrenamiento y R^2 de 0.909 en el conjunto de testeo. En este caso se pudo observar un leve sobreajuste (elevada varianza) del modelo a los datos que podría disminuirse utilizando menos estimadores

o cambiando el tipo de regularización empleada.

A pesar de que el modelo no establece una relación lineal entre las variables, sí permite hacer una apreciación de las features a partir de la cantidad la mejora marginal que se obtiene de la varianza explicada cuando se realiza una partición en el espacio de dicha variable exógena. En efecto, retorna un ranking en función de esa información aportada. En este caso, se observó (ver Gráfico 4.1) que las variables que mayor información aportan son las relacionadas a la mensura de la propiedad ($b - ratio$, $tsqm2$, $bsqm$, $tsqm$, $bsqm2$ y $bedbathratio$) así como también las variables de localización (lat y lon).

4.2.2 CATBOOST CON VARIABLES AUTOCORRELACIONADAS

Para evaluar si la creación de la matriz de adyacencia y la generación de las características ponderadas de los vecinos aportan valor en este tipo de modelos, se ajustó una nueva instancia del modelo donde lo único que cambió fue la inclusión de las variables lag . Nuevamente, se repitió el espacio multidimensional de hiper-parámetros para encontrar el mejor ajuste posible. En esta ocasión, se entrenaron árboles de decisión de profundidad máxima de 4 hojas, tasa de aprendizaje de 0.01 y 1000 estimadores. Con esta configuración se obtuvo un R^2 en entrenamiento de 0.928 y un R^2 en el conjunto de testeo de 0.893, lo que implicó una leve caída respecto del modelo anterior, con menor sobreajuste al mismo tiempo, pero que queda dentro de la variabilidad propia del algoritmo, que es no determinístico³.

³CatBoost FAQ, CatBoost, <https://catboost.ai/docs/concepts/faq.html>, (consultada el 15 de septiembre de 2020)

Respecto de la importancia relativa de las variables exógenas, se observó (ver Gráfico 4.3) una historia muy similar al modelo anterior, donde las mismas variables se encuentran entre las principales, excepto por el orden de importancia, lo cual sugiere que comparten algún grado de correlación entre las mismas. Respecto de las variables nuevas que se incluyeron en este modelo, las únicas que aparecieron ser relevantes fueron *lag_b_ratio*, *lag_bed_bath_ratio* y *lag_tsqm2*, pero este tipo de resultados puede ser explicado simplemente por la correlación existente entre las variables originales y las autocorrelacionadas.

4.3 Métodos de conjuntos simultáneos

4.3.1 RANDOM FOREST SIN VARIABLES AUTOCORRELACIONADAS

Luego de separar aleatoriamente el conjunto de entrenamiento y el de validación, se procedió a ajustar el algoritmo de RandomForest bajo un esquema de búsqueda de optimización de hiper-parámetros con validación cruzada de tres pliegues donde se obtuvieron como mejores resultados, los siguientes valores para el algoritmo: Profundidad máxima de cada árbol: sin límite. - Mínimo de muestras por hoja: 1. - Mínimo de muestras para realizar partición: 2. - Cantidad de estimadores: 50.

Al mismo tiempo, se buscó minimizar los errores cuadráticos como criterio de aprendizaje. Los resultados obtenidos fueron un R^2 de 0.985 en el conjunto de entrenamiento, 0.894 promedio de R^2 en validación cruzada

y 0.903 en el conjunto de datos de testeo.

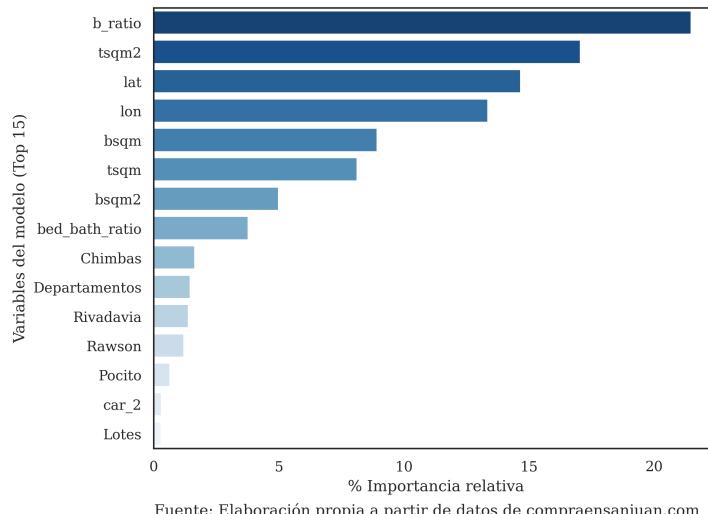
Debido a que tanto CatBoost como RandomForest están construidos a partir de árboles de regresión y clasificación, este último también computa la importancia relativa de las variables exógenas en función de la varianza explicada. En base a ese criterio, las características más relevantes para el modelo fueron sólo 5 variables: el ratio de metros construidos y metros totales (*b_ratio*), el tamaño del terreno y su transformación cuadrática (*tsqm* y *tsqm2*) y la latitud (*lat*) y longitud (*lon*).

4.3.2 RANDOM FOREST CON VARIABLES AUTOCORRELACIONADAS

A efectos de comprobar la misma hipótesis observada con el algoritmo de regresión de CatBoost, se optó por ajustar un modelo con las variables lag. Se procedió a repetir el espacio de posibilidades de hiper-parámetros y se obtuvieron los mismos resultados, excepto la cantidad total de estimadores que pasó de 50 a 500. Esto puede deberse a que al incrementar la cantidad de descriptores, se requiera de mayor varianza de los árboles para poder aportar más información.

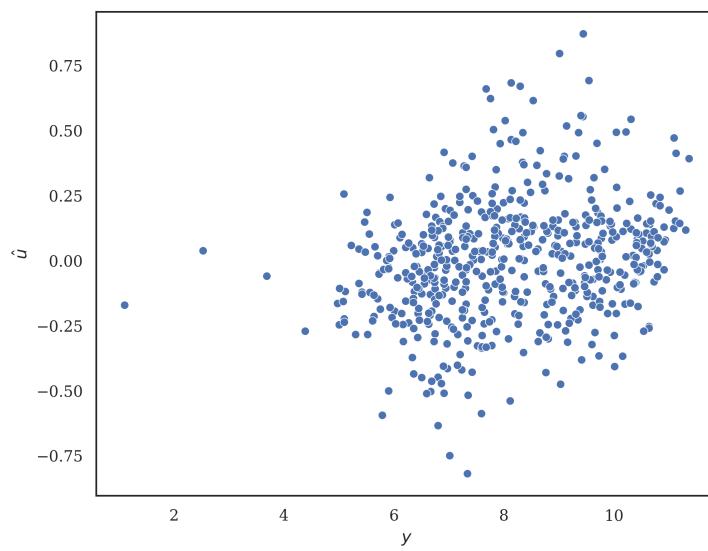
En cuanto a la performance del modelo, se observó un claro ejemplo de sobreajuste a los datos ya que se obtuvo un R^2 de entrenamiento de 0.981, un R^2 de validación cruzada de 0.857 y un R^2 en el set de testeo de 0.876. Los principales motivos para estos resultados yacen en el hecho de que se duplicó la cantidad de features al modelo y se aumentó considerablemente la cantidad de estimadores. Además, las variables lag que se agregaron están

fueramente correlacionadas con las originales, lo cual provoca aumento de la varianza de los estimadores. Finalmente, se pudo observar que, en cuanto a la importancia relativa de las variables exógenas, no cambió mucho el panorama, sólo que se incluyeron entre las principales algunas de las variables lag, tales como: *lag_b_ratio* y *lag_tsqm*. De todas formas no son efectos necesariamente relevantes debido a la alta correlación que existe entre las variables lag y las originales.



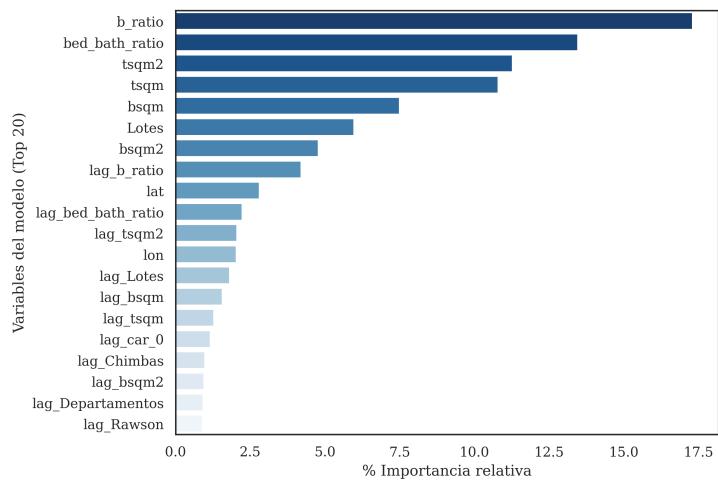
Fuente: Elaboración propia a partir de datos de compraensanjuan.com

Figura 4.1: Importancia relativa de las características según CatBoost, sin Lag X.



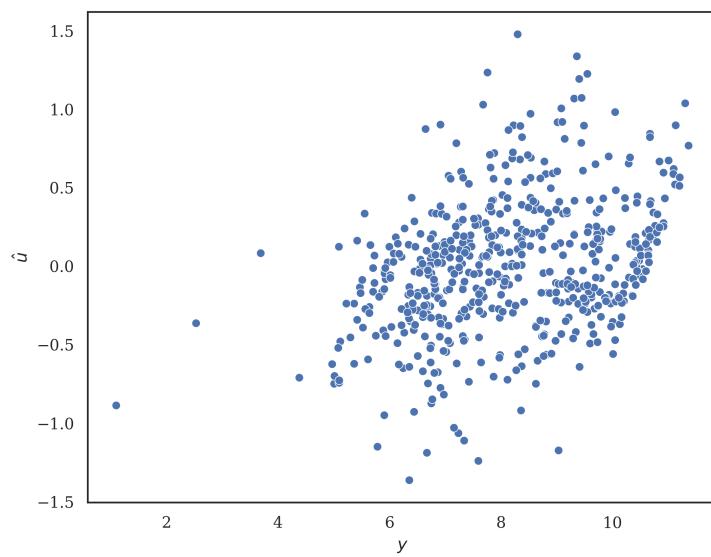
Fuente: Elaboración propia a partir de datos de compraensanjuan.com

Figura 4.2: Residuos de CatBoost, sin Lag X.



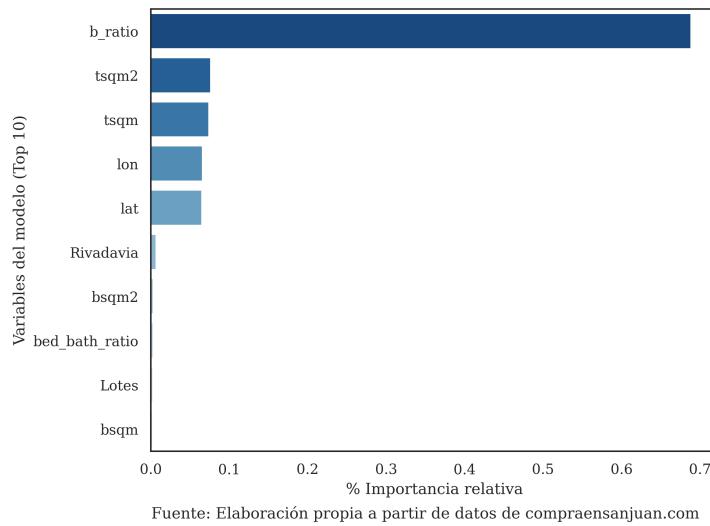
Fuente: Elaboración propia a partir de datos de compraensanjuan.com

Figura 4.3: Importancia relativa de las características según CatBoost, con Lag X.



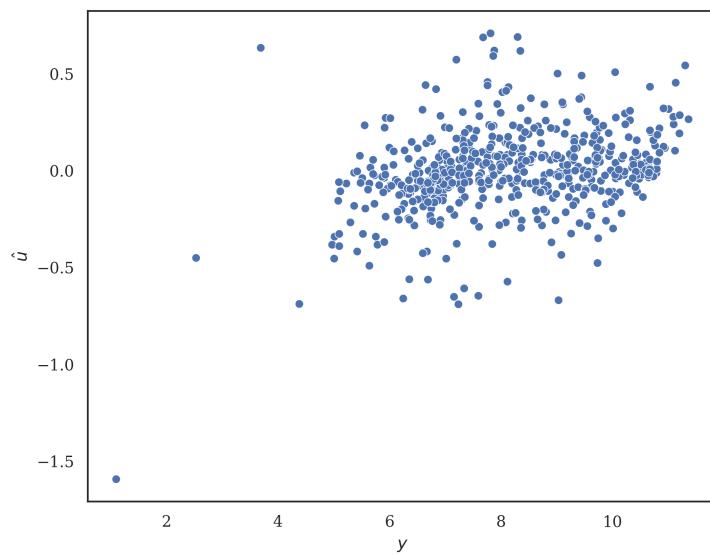
Fuente: Elaboración propia a partir de datos de compraensanjuan.com

Figura 4.4: Residuos de CatBoost, con Lag X.



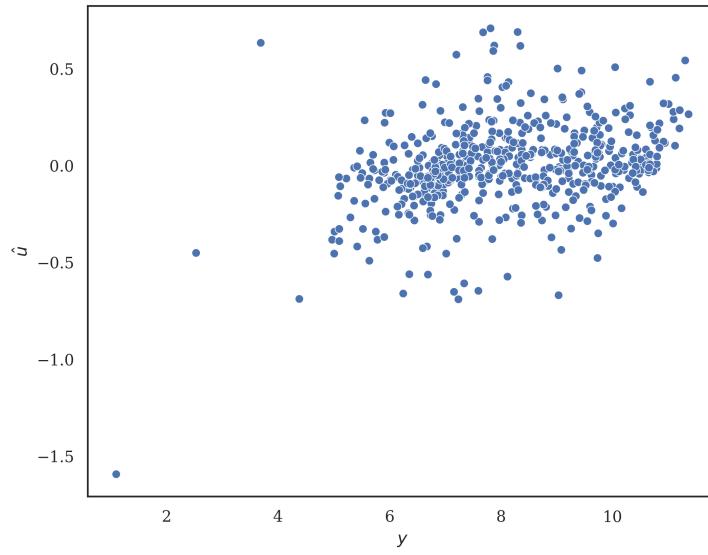
Fuente: Elaboración propia a partir de datos de compraensanjuan.com

Figura 4.5: Importancia relativa de las características según RandomForest, sin Lag X.



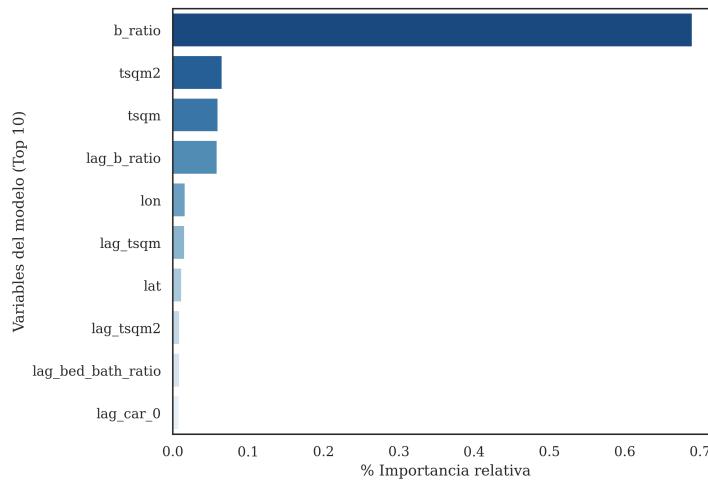
Fuente: Elaboración propia a partir de datos de compraensanjuan.com

Figura 4.6: Residuos de RandomForest, sin Lag X.



Fuente: Elaboración propia a partir de datos de compraensanjuan.com

Figura 4.7: Importancia relativa de las características según RandomForest, con Lag X.



Fuente: Elaboración propia a partir de datos de compraensanjuan.com

Figura 4.8: Residuos de RandomForest, con Lag X.

Conclusión

El foco de esta investigación estuvo en tratar de resolver una problemática aparentemente sencilla utilizando distintos métodos estadísticos a efectos de evaluar los puntos a favor y en contra de cada enfoque. Como se pudo exponer, existen distintas herramientas a disposición de los economistas a la hora de realizar modelos estadísticos en un área determinada, como es el caso de la predicción de precios inmobiliarios.

Si bien los distintos enfoques tuvieron resultados en cierta medida similares, existen ciertos matices que son necesarios de resaltar. El primero a remarcar y sobre el cual no se hizo mención en toda la investigación hasta el momento, es tener en claro cuál es el objetivo del investigador. Si la intención radica en tratar de explicar la relación entre las variables exógenas y la variable objetivo, lo ideal es aplicar modelos lineales, con las transformaciones pertinentes y teniendo todos los recados necesarios para presentar e interpretar los valores, a efectos de lograr resultados representativos, insesgados y robustos. Una extensión de este enfoque consiste en aplicar conocimientos específicos de la problemática planteada -en este caso, nociones de dependencia espacial- con el objetivo de lograr una descomposición más profunda de la problemática.

Por el contrario, si las intenciones yacen en tratar de lograr la mejor predicción posible, estos requisitos formales y elementos a tener en cuenta a la hora de especificar el problema se pueden evitar utilizando modelos no paramétricos. Como se pudo demostrar, el gran poder predictivo de los métodos de conjuntos quedó demostrado a pesar de que haya existencia de dependencia espacial, efecto que en una regresión lineal sencilla puede ocasionar el cálculo de estimadores inefficientes o sesgados. Sin lugar a dudas, no todo lo que reluce es oro y, en el caso de los modelos de conjuntos, la pérdida de interpretabilidad se hace notar. Con estos modelos, si bien es posible evitar algunos dolores de cabeza y el tener que pensar un poco más activamente los problemas, también limitan la posibilidad de analizar los motivos por los cuales el modelo puede estar performando mal, así como también nos quita noción de las interacciones existentes entre las variables que pueden aportar valor al análisis.

Para concluir, y con intenciones de responder a la pregunta planteadas por esta investigación, existe un universo de algoritmos que quedan fuera del alcance de los programas universitarios de las carreras de economía que tienen grandes posibilidades de aplicación en la vida cotidiana de los profesionales, tanto economistas como otros profesionales. Es absolutamente necesario tomar estos algoritmos con pinzas para evitar encontrar relaciones espúreas en base a los datos disponibles, por lo que es altamente recomendable investigar y conocer sobre la problemática subyacente antes de implementarlos.

Referencias

- Anselin, L., 1988. *Spatial Econometrics: Methods and Models*, Springer, Dordrecht. Available at: <https://link.springer.com/book/10.1007/978-94-015-7799-1>.
- Bartik, T.J., 1987. The Estimation of Demand Parameters in Hedonic Price Models. *Journal of Political Economy*, 95(1), pp.81–88. Available at: <http://www.jstor.org/stable/1831300>.
- Basu, Sabyasachi & Thibodeau, Thomas G, 1998. Analysis of Spatial Autocorrelation in House Prices. *The Journal of Real Estate Finance and Economics*, 17(1), pp.61–85. Available at: <https://ideas.repec.org/a/kap/jrefec/v17y1998i1p61-85.html>.
- Bourassa, Steven, Hoesli, Martin & Peng, Vincent, 2002. Do Housing Submarkets Really Matter? *. *Journal of Housing Economics*, 12, pp.12–28.
- Brian J. L. Berry & Robert S. Bednarz, 1975. A Hedonic Model of Prices and Assessments for Single-Family Homes: Does the Assessor Follow the Market or the Market Follow the Assessor? *Land Economics*, 51(1), pp.21–40. Available at: <http://www.jstor.org/stable/3145138>.
- Dubin, R.A., 1988. Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms. *The Review of Economics and Statistics*, 70(3), pp.466–474. Available at: <http://www.jstor.org/stable/1926785>.
- Elizabeth C. Hirschman & Morris B. Holbrook, 1982. Hedonic Consumption: Emerging Concepts, Methods and Propositions. *Journal of Marketing*, 46(3), pp.92–101. Available at: <http://www.jstor.org/stable/1251707>.
- Herath, Shanaka & Maier, Gunther, 2010. *The hedonic price method in real estate and housing market research. A review of the literature*, WU Vienna University of Economics; Business. Available at: <https://ideas.repec.org/p/wiwiw/wus009/588.html>.
- Hill, R., 2011. Hedonic Price Indexes for Housing. Available at: <https://www.oecd-ilibrary.org/content/paper/5kghzxpt6g6f-en>.
- Jackson, J.R., 1979. Intraurban variation in the price of housing. *Journal of Urban Economics*, 6(4), pp.464–479. Available at: <https://ideas.repec.org/a/eee/juecon/v6y1979i4p464-479.html>.
- Keith Cowling & John Cubbin, 1972. Hedonic Price Indexes for United Kingdom Cars. *The Economic Journal*, 82(327), pp.963–978. Available at: <http://www.jstor.org/stable/2230261>.

- Lancaster, K.J., 1966. A New Approach to Consumer Theory. *Journal of Political Economy*, 74(2), pp.132–157. Available at: <http://www.jstor.org/stable/1828835>.
- Lesage, J., 1999. The theory and practice of spatial econometrics.
- Malpezzi, S., 2003. *Hedonic pricing models: A selective and applied review*, University of Wisconsin Center for Urban Land Economic Research. Available at: <https://EconPapers.repec.org/RePEc:wop:wisule:02-05>.
- Mitchell, T.M., 1997. *Machine Learning*, McGraw-Hill. Available at: <https://books.google.com.ar/books?id=EoYBngEACAAJ>.
- Moulton, B.R., 1996. Bias in the Consumer Price Index: What Is the Evidence? *Journal of Economic Perspectives*, 10(4), pp.159–177. Available at: <https://www.aeaweb.org/articles?id=10.1257/jep.10.4.159>.
- Pace, R Kelley & Gilley, Otis W, 1997. Using the Spatial Configuration of the Data to Improve Estimation. *The Journal of Real Estate Finance and Economics*, 14(3), pp.333–340. Available at: <https://ideas.repec.org/a/kap/jrefec/v14y1997i3p333-40.html>.
- Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, pp.2825–2830.
- Peter F. Colwell & Gene Dilmore, 1999. Who was first? An examination of an early hedonic study. *Land Economics*, 75(4), pp.620–626. Available at: <http://www.jstor.org/stable/3147070>.
- Prokhorenkova, Liudmila et al., 2018. CatBoost: unbiased boosting with categorical features. In S. Bengio et al., eds. *Advances in neural information processing systems*. Curran Associates, Inc. Available at: <https://proceedings.neurips.cc/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf>.
- Rosen, S., 1974. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, 82(1), pp.34–55. Available at: <http://www.jstor.org/stable/1830899>.
- Trevor Hastie, Robert Tibshirani & Jerome Friedman, 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, NY.