Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración Pública

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL DE ESPECIALIZACIÓN

Modelos de valuación para el sector inmobiliario de CABA

Desarrollo con métodos analíticos predictivos

AUTOR: MARIA LAURA LORENZO
MENTOR: ROBERTO ABALDE

SEPTIEMBRE 2022

Resumen

Este trabajo se propondrá establecer los lineamientos para el aprovechamiento de los grandes volúmenes y tipos de datos disponibles en el sector inmobiliario aplicando *big data analytics* y *machine learning* a la predicción de precio de propiedades en venta en la Ciudad Autónoma de Buenos Aires.

El objetivo del proyecto es poder estimar los valores de propiedades de Capital Federal, analizando cuales son los determinantes del precio de los departamentos. Se intentará también poder ofrecerle al mercado inmobiliario una referencia de precios para la toma decisiones, mejorando los tiempos de las operaciones inmobiliarias. Esto se logrará entendiendo mejor la oferta y optimizando el tiempo entre la voluntad de venta y el hecho en sí, haciendo más rentable la operatoria de la organización.

En cuanto al desarrollo, se utilizarán herramientas estadísticas en conjunto con los valores del mercado. Se tomarán aquellos factores de la propiedad que afecten a su valor, como lo son cantidad de ambientes, barrio donde está ubicado, tipo de propiedad, cercanía de medios de transporte y otros posibles elementos que puedan tener relación con el valor de la propiedad.

Palabras clave: Big Data, Business Analytics, Mercado Inmobiliario, Predicción de Precios, machine learning

Tabla de contenido

Introducción

La disponibilidad de datos y la capacidad de construir conocimiento presenta una oferta cada vez más amplia. Pero ¿cómo pueden estos datos asistir en lo que las organizaciones resuelven cada día? La optimización en la toma de decisiones ha sido objeto de estudio desde hace más de medio siglo y el término *business intelligence*, como método basado en datos para lograr la optimización, fue popularizado hace nada más y nada menos que 33 años (Power D. , 2007).

Brynjolfsson, Hit y Kim (2011) encontraron que las compañías que basan sus decisiones en datos son más productivas y mostraron una correlación entre este tipo de práctica al momento de tomar decisiones y un aumento en la rentabilidad, en el valor de mercado, y una optimización en la utilización de bienes dentro de estas compañías.

Tras la experiencia de implementación de soluciones de *business intelligence*, se han creado marcos de trabajo que reúnen los factores críticos de éxito con el fin de asistir a las empresas al momento de emprender proyectos de este tipo.

Sin embargo, año tras año los datos digitales existentes a nivel mundial crecen de forma exponencial llegando a niveles nunca antes vistos. Este volumen ya no puede ser manejado utilizando bases de datos relacionales. Los entornos tradicionales de datos y los métodos de *business intelligence* presentan limitaciones para manipular múltiples formatos y grandes volúmenes. Los tiempos de procesamiento se vuelven insostenibles. Esta situación ha llevado a la creación de nuevas soluciones que dan origen al concepto de *big data analytics*.

Este trabajo se propone establecer lineamientos necesarios para el aprovechamiento de los grandes volúmenes y tipos de datos disponibles, que se encuentran en constante crecimiento, mediante la implementación de *big data analytics* en el área del mercado inmobiliario.

El objetivo es el desarrollo de modelos de aprendizaje automático para el mercado inmobiliario de la Ciudad Autónoma de Buenos Aires, siendo que no se ha encontrado una

herramienta para la cotización inmediata de propiedades en base a distintas variables características de un inmueble.

El recorrido a seguir para alcanzar el objetivo general de este estudio se basará en los siguientes objetivos específicos:

- Definir la relación existente entre *business intelligence* y *big data analytics*
- Adaptar el marco de trabajo al mercado inmobiliario.
- Aplicación de métodos analíticos predictivos aplicados a *business intelligence*

El beneficio que se espera obtener de este trabajo es poder ayudar al sector inmobiliario a cotizar de manera inmediata nuevas propiedades a la venta para que puedan ofrecer un mejor servicio a sus clientes y optimizando recursos al ahorrar tiempo y dinero para la concreción de ventas.

Davenport (2009) dice que una de las formas de mejorar la toma de decisiones es refinando el análisis de datos. Además, destaca que las organizaciones que utilizan el análisis de datos como ventaja competitiva aplican los datos de manera de optimizar operaciones en grados sin precedencia y transforman la tecnología de una herramienta de soporte en un arma estratégica (Davenport, 2006). Por otro lado, resalta:

En un momento en que las compañías en muchas industrias ofrecen productos similares y utilizan tecnologías comparables, los procesos de negocio son los puntos de diferenciación fundamentales. Y las organizaciones que compiten utilizando el análisis de datos extraen cada gota de valor de esos procesos. Entonces, como otras compañías, ellas saben qué productos sus clientes quieren, pero también saben qué precios sus clientes pagaran, cuántos productos cada cliente comprará en su vida y qué desencadena que el público compre más (Davenport, 2006).

Weill y Aral (2006) realizaron un estudio en el cual recopilaron información sobre inversiones de Tecnología de la Información (TI) de 147 compañías norteamericanas en un periodo de 5 años y encontraron, entre otros resultados, que las organizaciones de mayor desempeño en la industria de ventas por mayor y menor y en la industria de transporte gastan 11% más en TI que el promedio, con portfolios que se orientan a la inversión para la obtención de información, indicando que la ventaja competitiva está en el uso efectivo de la información.

El presente trabajo está organizado en cinco partes: Introducción, Fundamentos, Base de datos, Aplicación de métodos analíticos predictivos y Conclusión. En la primera parte se ha presentado la introducción, que se centró en la definición del problema y en la determinación del objetivo general y de los objetivos específicos que dirigen esta investigación.

La segunda parte describirá el marco teórico donde se definen los conceptos generales que incluyen *business intelligence, business analytics, big data y big data analytics*. Además se estudiarán distintos marcos de trabajo para la implementación de *business intelligence* y se ubica *business intelligence* en el contexto del mercado inmobiliario de la Ciudad Autónoma de Buenos Aires.

En la tercera parte se analizará la base de datos obtenida para el análisis, con la limpieza de la información y un análisis de estadística descriptiva del mismo. En la cuarta parte, se propondrán distintos modelos y se seleccionará el óptimo para el armado de la herramienta para el mercado inmobiliario. Finalmente, se presentarán las conclusiones obtenidas como resultado de esta investigación.

Fundamentos

La tecnología cambia la forma de hacer los negocios, y el sector inmobiliario no es ajeno a esta evolución. El mercado inmobiliario actual de la Ciudad Autónoma de Buenos Aires es muy heterogéneo, tanto en aspectos constructivos como en las características barriales. Es de esperar que esta heterogeneidad se vea reflejada en los precios de oferta de los departamentos cuando éstos salen a la venta. Más allá del conocimiento que tenga cada propietario o cada agente inmobiliario de la propiedad a tasar y del entorno del lugar en el que se localiza la misma para poder asignar un precio de oferta, resulta interesante conocer técnicamente cuáles son los determinantes más relevantes en la formación de dicho precio.

Transformar los datos en información para generar conocimiento y optimizar la toma de decisiones inteligentes en los negocios es una herramienta fundamental y muy valiosa. Los datos son la puerta de entrada para el desarrollo de productos y servicios. En el caso del mercado inmobiliario, el análisis de datos permite, entre otras cosas, categorizar a las propiedades e identificar cuál es el precio óptimo al que puede ofertarse.

En el sector inmobiliario, hasta hace relativamente poco, no se confería ninguna importancia a la recopilación de datos de los clientes. Simplemente, se acumulaban en bases de datos de inmobiliarias, sin que todo ese material tuviera un destino claro fuera de los resúmenes anuales de ventas (InmoGesco, 2022).

Pero el *Big Data* para una inmobiliaria no es sólo recopilación de datos. Es, sobre todo, análisis de estos. De ahí han surgido conceptos como el análisis predictivo que se basa en adelantarse a la competencia con la identificación de potenciales oportunidades de negocio. Un análisis de este tipo podría localizar e identificar las casas y los propietarios que estarían en disposición de vender, para así adelantarse a la competencia.

Las inmobiliarias con negocios puramente digitales (que serán llamadas *PropTech* en este trabajo), simplifican el proceso de compra y venta, volviéndolo una experiencia sencilla,

rápida y satisfactoria para el consumidor. Dentro del movimiento *PropTech* una de sus ramificaciones más comunes y destacadas es la del *Big Data* inmobiliario. Tradicionalmente, las empresas del sector han fundamentado sus decisiones en una combinación de intuición e información de corte tradicional. Sin embargo, hoy, como bien explica (Gamarra, 2020) "es posible combinar grandes bases de datos para predecir el precio por metro cuadrado de un alquiler de tres años en la ciudad de Seattle, por poner un ejemplo". Para mayor abundancia, la cantidad de datos que genera esta industria y sus usuarios hace presuponer que las empresas que se especialicen en el *Big Data* no solo no tengan ningún problema a la hora de adentrarse en el mercado, sino que disfrutarán de una muy relevante ventaja competitiva sobre empresas ya establecidas (Deloitte, 2018). Comprender y actuar en función del estado de tal cuestión se convierte en una decisión crucial para cualquier empresa o usuario inmerso en el mercado inmobiliario.

Un sistema con *big data y machine learning* puede captar problemas con los datos y al mismo tiempo resolverlos. (Lescano, 2019). Las técnicas de *Machine Learning* detectan patrones de comportamiento que permiten conocer al consumidor, facilitando la compraventa de propiedades. Las herramientas de *Machine Learning* también aportan datos de entrenamiento para detectar el ciclo de vida de cada propiedad *(Lifetime Value)* (Paul D. Berger, 1998), y, a partir de ello, actuar en consecuencia, sobre qué estrategia seguir para concretar la venta sabiendo qué oferta hacer.

Ante el vértigo que todo un proceso de transformación digital podría ocasionar, el uso de las técnicas de aprendizaje automático son un buen primer paso. Apoyarse en técnicas de predicción para estimar la valoración de las propiedades es un factor clave y diferenciador. Estas técnicas dan soporte a las decisiones que deben tomar los vendedores de dos formas distintas. Por un lado, generando una oferta de manera veloz para la propiedad en cuestión que será atractiva para el mercado, lo que ahorrará el tiempo de espera entre la oferta y la efectiva venta de la propiedad, a la vez que se reduce el costo asociado a esa espera. Por el otro, generando expectativas reales en cuanto al rédito que se puede obtener por la venta o compra de una propiedad, fundamental para la toma de decisiones.

Como resultado, las inmobiliarias (o vendedores particulares) podrían reducir costos y mejorar la efectividad del proceso de compra/venta, ofreciendo precios adecuados al mercado. Contar con la infraestructura adecuada para capturar y procesar la información se vuelve vital para acoplarse a la marcha tecnológica actual.

Big Data* y su relación con *Business intelligence

Business intelligence ayuda a recopilar información esencial de una amplia variedad de datos no estructurados y los convierte en información procesable que permite a las empresas tomar decisiones en base a información y mejorar la eficiencia y la productividad comercial. Los desafíos que enfrenta cualquier organización en la inteligencia empresarial y la toma de decisiones incluyen la falla del plan, la falta de preparación, la falla de los recursos y la capacidad de asumir riesgos (Yanfang Niu, 2021).

El *Big Data* y el *Business Intelligence* son dos tecnologías que deben ser conocidas por cualquier empresa que vaya a iniciar un proceso de cambio. Durante el año 2019 y según la consultora Gartner, la primera prioridad de inversión para las empresas inmersas en procesos de transformación digital fue la analítica de datos (43%), seguida por la ciberseguridad (43%) y las soluciones y servicios *Cloud Computing* (39%), es decir, necesidades, sobre el dato que hace que los nuevos líderes digitales precisen de conocimientos y competencias adicionales sobre estas tecnologías.

El *Big Data* y el *Business Intelligence*, por sus similitudes, generan mucha confusión entre empresarios, emprendedores y directivos. El concepto *Big Data* hace referencia a un conjunto de tecnologías y herramientas capaz de capturar, almacenar y procesar grandes cantidades de datos en tiempo y coste asumibles para una organización. El *Business Intelligence* o inteligencia de negocio consiste en un conjunto de técnicas de gestión empresarial que permiten a una organización tomar decisiones de negocio en base a datos, que han sido tratados por distintas herramientas para convertirlos en información.

Por tanto, en cuanto a sus diferencias: mientras el *Big Data* se centra en la captura, almacenamiento y procesamiento de los datos, el *Business Intelligence* se centra en los procesos

de análisis de dichos datos para convertirlos en información y tomar las decisiones de negocio oportunas.

Se puede decir que el dato siempre ha sido la fuente para procesos de *Business Intelligence.* Lo que ha cambiado en la actualidad es que el dato se ha vuelto masivo. Además, se le ha dado un nombre a las tecnologías capaces de tratarlo: *Big Data*. El perfil de las personas que trabajan directamente con cada una de éstas tecnologías también es diferente. Así, mientras que en los equipos de trabajo relacionados con el *Big Data* aparecen perfiles como ingenieros, estadísticos y matemáticos, los equipos de trabajo de *Business Intelligence* de una empresa están formados por expertos en administración de empresas, economistas, expertos en marketing y, de nuevo, ingenieros y técnicos. Estos no son los perfiles que habitualmente se encuentran en el mercado inmobiliario.

Lo que sí es cierto es que tanto *Big Data* como *Business Intelligence* son complementarios y ayudan a las organizaciones a lograr una visión 360 del negocio y, consecuentemente, a tomar decisiones acertadas en base a unos datos. Además, cuando estos se transforman en conocimiento, les permiten ofrecer servicios y experiencias innovadores al cliente, surgiendo un modelo de negocio de éxito. (https://blog.powerdata.es, 2021). Los buenos resultados obtenidos por los *Big Business* llevaron a ejecutivos de todos los sectores a redescubrir el poder potencial de la información. Por ello, la inversión en *Big Data* y *Business Intelligence* es sólo el principio para lograr una cultura empresarial basada en datos. Con el enfoque correcto, hoy es posible utilizar grandes datos para formar ideas, transformar las operaciones de negocios y mejorar las experiencias de los clientes.

La optimización en la toma de decisiones ha sido objeto de estudio desde la década de los sesenta, con la implementación de los primeros sistemas de soporte de decisión. (Power D. J., A brief history of decision support systems, 2007). En su investigación sobre la historia de los sistemas de soporte de decisión, Power (2007) relata que el término *Business Intelligence* es popularizado recién en 1989 por un analista de Gartner[1] llamado Howard Dresner, y la describe como:

[1] Gartner es una organización dedicada a la investigación de las tecnologías de la información, fundada en 1979 con sede en Stamford, Connecticut, USA. Provee consultoría y programas ejecutivos entre otros servicios.

"Un conjunto de conceptos y métodos para mejorar la toma de decisiones utilizando sistemas de soporte basados en información. Los términos *Business Intelligence* y compendios de información, reportes, herramientas de consulta y sistemas ejecutivos de información suelen utilizarse indistintamente. En general los sistemas de BI son sistemas de soporte de decisión basados en datos" (Power, 2007).

Por otro lado, *The Data Warehousing Institute* (TDWI)[2], destaca:

Business Intelligence unifica datos, tecnología, análisis y conocimiento humano para optimizar las necesidades del negocio y en última instancia conducir el éxito de la empresa. Los programas de inteligencia de negocios generalmente combinan un *data warehouse* empresarial y una plataforma de *Business Intelligence* o un conjunto de herramientas para transformar datos en información comercial que pueda ser utilizada y que dirija un plan de acción (TDWI, s.f.).

Marcos de trabajo para la implementación

La transformación de datos en información que provee *Business Intelligence* lleva a distinguir dos tipos de sistemas: los sistemas operacionales y los informacionales. *On-line transaction processing* (OLTP) se refiere a los sistemas que se utilizan en las operaciones diarias de una empresa, que facilitan el manejo de las aplicaciones basadas en transacciones, son sistemas de tipo operacionales. OLTP permite crear, actualizar y borrar registros como por ejemplo en un sistema de órdenes de compra o de transacciones financieras. Estas tareas se basan generalmente en sistemas de bases de datos relacionales. Algunas de sus características son tiempo de respuesta corto, alta concurrencia y alta disponibilidad (Oracle, 2011)

Por otro lado, *on-line analytical processing* (OLAP) permite el análisis de datos transformándolos en un reflejo de las dimensiones reales de la compañía para la toma de decisiones, son sistemas de tipo informacionales. Estas tareas están basadas generalmente en un *data warehouse*, que es el utilizado por los programas de *Business Intelligence*.

[2] TDWI es el principal instituto de educación en el área de BI y Data Warehousing, fundado en 1995. Se dedica a la investigación, educación y certificación de profesionales dentro del sector de tecnología de la información.

La diferencia entre ambos se resume en la siguiente tabla:

Tabla 1: Diferencias entre sistemas OLTP y OLAP

	Operacional (OLTP)	***Informacional (OLAP)***
Contenido de datos	Valores actuales	Archivados, derivados, resumidos
Estructura de datos	Optimizada para transacciones	Optimizada para consultas complejas
Frecuencia de acceso	Alta	Mediana a baja
Tipo de acceso	Lectura, actualización, borrado	Solo lectura
Utilización	Predecible, repetitiva	Aleatoria, para situaciones específicas, para descubrimiento
Tiempo de respuesta	Milisegundos	Segundos a minutos
Usuarios	Muchos	Pocos

Fuente: Adaptado de Ponniah, P. (2004). Data warehousing fundamentals: a comprehensive guide for IT professionals. John Wiley & Sons.

Adicionalmente a esto, toda iniciativa de *Business Intelligence* requiere de una infraestructura: el *data warehouse.* Según Inmon (2005), la definición dice que "un *data warehouse* es una colección de datos orientada a un tema, integrada, no volátil y variante en el tiempo, utilizada para avalar las decisiones de la gerencia. El *data warehouse* contiene datos corporativos granulares". En contraste con los sistemas que se utilizan en las operaciones diarias de una empresa, el *data warehouse* almacena datos orientados a un tema. Por ejemplo, un negocio de ventas puede organizar su información para procesar órdenes de compra, facturar a sus clientes, ordenar cuentas a cobrar y cuentas a pagar, todos procesos que soportan sus actividades del día a día. Pero su *data warehouse* estaría organizada por tema, de acuerdo a las prioridades del negocio, por ejemplo, ventas, productos, clientes, etc.

Los datos son integrados porque provienen de distintas aplicaciones que soportan los procesos del negocio. Para integrar los datos es necesario eliminar las inconsistencias que

puedan existir entre las distintas aplicaciones y fijar estándares. Continuando con el ejemplo anterior, para el tema cliente se obtendrían datos de los sistemas de órdenes de compra, de facturación, de cuentas a cobrar, etc., relacionadas con ese cliente en particular. La integración implica estandarizar códigos, medidas, convenciones para nombres, etc. El *data warehouse* es no volátil porque almacena los datos en masa y no se actualiza con cada transacción que realiza la empresa. Por ejemplo, la aplicación de órdenes de compra debe permitir leer, insertar, editar y borrar cada transacción mientras que el *data warehouse* solo permitirá leer los datos de las órdenes de compra para un determinado momento.

La característica de variante en el tiempo se relaciona con la manera de almacenar los datos: se crea una foto que representa un período de tiempo. Por ejemplo, la base de datos de las cuentas a cobrar mantiene la información actualizada con cada transacción mientras que el *data warehouse* contiene una secuencia histórica de cada registro de las cuentas a cobrar por cliente para el periodo de tiempo estipulado en esa foto. La granularidad se refiere al nivel de detalle: a mayor nivel de detalle, mayor granularidad. Por ejemplo, se podría evaluar el resumen de las compras realizadas por un cliente en los últimos seis meses o cada compra realizada por un cliente en los últimos seis meses, incrementando de esta manera el nivel de detalle. Como este enfoque alcanza a toda la organización se lo conoce como *Enterprise data warehouse*. Existen otros enfoques que se basan en la creación de *data marts*, cuyo alcance se limita a un departamento de la empresa (Inmon, 2005)

La estructura de datos de un *data warehouse* difiere de la de una base de datos relacional. Las dos dimensiones de una tabla en el modelo relacional se extienden a un modelo multidimensional. La arquitectura del *data warehouse* está dividida en tres áreas funcionales: adquisición de datos, almacenamiento y presentación de la información. Los datos se obtienen de distintas fuentes operacionales y son limpiados y transformados antes de ser cargados en el *data warehouse*. (Ponniah, 2004)

La adquisición y el almacenamiento de los datos está soportado por tres procesos de soporte que se conocen como ETL, *extract-transform and load* o extraer-transformar y cargar. Estos procesos incluyen la extracción de datos de los distintos sistemas fuentes, su transformación a los formatos y estructuras apropiadas para su almacenamiento en la base de

datos del *data warehouse* y el traslado de los datos al repositorio del mismo. La carga de los datos es una función que requiere tiempo, por ello el negocio deberá evaluar los ciclos de recarga según sus necesidades (Ponniah, 2004).

La presentación de la información incluye diferentes métodos que se adaptan a las necesidades de los distintos tipos de usuarios, desde los más novatos hasta los más avanzados. Las consultas *ad hoc* permiten al usuario definir la información que necesita y componer sus propios *queries*, resultando en reportes *online*. Existen además reportes con formatos predefinidos donde el usuario ingresa los parámetros necesarios; el reporte puede ser programado para un determinado horario o bien ser ejecutado cuando el usuario lo requiera (Ponniah, 2004).

Por otro lado, existen representaciones gráficas de la información de negocio llamadas visualizaciones, que pretenden mejorar la comprensión del significado de los datos, exponiendo patrones, tendencias y correlaciones que podrían no ser detectados en un texto. Las visualizaciones son especialmente útiles cuando se manejan grandes cantidades de información (Rouse, 2010)

Business intelligence **en el mercado inmobiliario**

Los modelos predictivos son patrones estadísticos de comportamiento futuro basados en factores variables que puedan influenciar los resultados, es decir se ocupan de pronósticos y tendencias. Los modelos predictivos se utilizan entre otras aplicaciones, para mejorar la experiencia de compra de los clientes, para planificar la capacidad de producción y para el manejo del cambio (Rouse, 2010).

El análisis multivariante implica la observación y el análisis de más de una variable al mismo tiempo, permitiendo el estudio de los datos a través de las distintas dimensiones, considerando los efectos de todas las variables con respecto a un hecho (Rouse, 2010).

Con respecto a las herramientas y capacidades, se proponen cinco categorías: descriptivas, diagnósticas, predictivas, prescriptivas y cognitivas. Las herramientas descriptivas

permiten explicar qué pasó y lo reflejan en *dashboards* y reportes. Las herramientas diagnósticas buscan distinguir el por qué y se representan en visualizaciones. Las predictivas, en cambio, se basan en análisis estadístico, *data mining* y modelos predictivos. Las herramientas prescriptivas indican la acción a seguir. Finalmente, las herramientas cognitivas integran todas las herramientas anteriores y utilizan correlaciones e hipótesis además de recordar y aprender de cada evento.

Por otro lado, *SAS Institute Inc* en 2018 (SAS Institute) sugiere ocho niveles de capacidades analíticas (herramientas): reportes *standard*, reportes *ad hoc*, consultas *drilldown* / OLAP, alertas, análisis estadístico, pronósticos, modelos predictivos y optimización. Los cuatro primeros niveles se basan en el reporte de datos históricos mientras que los últimos cuatro se orientan hacia una visión predictiva y otorgan respuestas a preguntas más complejas.

En la tabla siguiente se presenta un cuadro comparativo entre las capacidades propuestas:

Tabla 2: Comparación de capacidades y herramientas

Capacidades	***Preguntas que responde***	***Reportes o herramientas***
Descriptivas	¿Qué pasó? ¿Cuándo pasó? ¿Cuánto? ¿Con qué frecuencia? ¿Dónde?	Reportes *standard* Reportes *ad hoc*
Diagnósticas	¿Por qué? ¿Dónde está el problema? ¿Cómo encuentro las respuestas?	Consultas *drilldown,* OLAP
Prescriptivas	¿Qué acción se debe tomar? ¿Cuándo reaccionar? ¿Qué acciones son necesarias ahora?	Alertas
Predictivas	¿Qué va a pasar?	Análisis estadístico Pronósticos

	¿Cómo sigue?	Modelos predictivos
	¿Cómo se verá afectado el negocio?	
Cognitivas	¿Qué va a pasar y cómo actuar basado en lecciones aprendidas?	Optimización

Fuente 1: Elaboración propia

La transformación para las inmobiliarias pasa por asumir los desafíos digitales y adoptar nuevas tecnologías. Son cambios que afectan a la comercialización de productos y servicios. Multitud de expertos reconocen que dentro de cada organización los departamentos de ventas representan uno de los principales nichos sobre los que focalizar las estrategias de digitalización y la aplicación de tecnología. Conseguir nuevos métodos de prospección de clientes para poder contactar con el cliente potencial en el momento oportuno es el sueño de toda empresa, pyme o autónomo. Y este sueño se vuelve cada vez más real gracias al avance del *Big Data* aplicado al *Business Intelligence*.

Un software de *business intelligence* proporciona un análisis en tiempo real de datos cruciales que permiten exactamente esto. Son de millones de datos externos alojados en Internet, que varían a diario, y que pueden cruzarse con la estrategia de manejo de clientes propia, dando como resultado información clave para penetrar en cualquier negocio.

Proyectos de nuevas obras, licencias ambientales, enajenaciones y ventas de terrenos y solares, información sobre edificios y propietarios, planes de inversión, informaciones complementarias como noticias publicadas en medios y datos de contacto son algunos tipos de segmentación. Porque en el sector inmobiliario, donde el mercado es cada vez más complejo y la competencia crece, tener información actualizada es clave para llegar antes que los demás.

Por ejemplo, la construcción de un supermercado, centro comercial o edificio de oficinas puede propiciar la aparición de nuevas viviendas, así como saber que una empresa está preparando su expansión nos dice que necesita un nuevo local de oficinas. Conocer los planes medioambientales y los planteamientos parcelarios de un municipio ayudarán a la planificación y gestión de proyectos inmobiliarios, e identificar los proyectos urbanísticos tanto públicos como privados desde sus inicios supone una palanca de crecimiento comercial.

Disponer en tiempo real de toda esta información hace que los comerciales conozcan con bastante profundidad el sector del cliente y puedan establecer un diálogo de calidad donde adapten su oferta. Porque a un cliente potencial le puede interesar la propiedad, pero sobre todo le interesa saber cómo lo que se le está ofreciendo cumple con la mayor parte de sus necesidades.

Definitivamente, el *Business Intelligence* se convierte en el mejor socio para multiplicar exponencialmente las oportunidades de negocio para cualquier inmobiliaria, independientemente de su tamaño.

Gestión de datos en contexto inmobiliario

Se puede entender que los datos capturados por la organización son su principal activo para la toma de decisiones. En este sentido, los datos solían estar representados por datos estructurados. El desarrollo de la tecnología de la información y la comunicación produjo un crecimiento exponencial de los datos que acumula la empresa. Si una empresa adicionalmente transitó varias fusiones y adquisiciones, la colección de datos también se encuentra disociada entre distintas formas en que se almacena la información.

Pese a que las inmobiliarias en la actualidad tienen un sistema de información centralizado donde cada uno de sus operadores (vendedores) comparte la información colectada sobre una nueva propiedad a la venta, este sistema es operador-dependiente. Es común encontrar casos donde se completa la información de manera desorganizada solo con el propósito que el sistema les permita avanzar a la pantalla siguiente sin respeto alguno por la información que la plataforma solicita. Habitualmente se encuentran casos donde el grueso de la información se concentra en el campo de texto donde se espera una breve descripción, incorporando ahí datos claves de la propiedad como ser características edilicias, ubicación, precio requerido, etc.

El conjunto de esta información es *Big data* y se entiende habitualmente como grandes volúmenes de datos generados y puestos disponibles en el ambiente digital actual. Boyd y Crawford (2012) entendían que sólo el volumen de los datos jamás hubiera sido suficiente para

encapsular la novedad del fenómeno de Big Data en las organizaciones. Se asocia con los grandes volúmenes de datos a la diversidad de la información, la frecuencia con la que es actualizada y, más generalmente, la velocidad a la que crece (O'Reilly, 2012; Davenport, 2014). Todos juntos, estos atributos pretenden elevan grandes, difusas y cambiantes estructuras de datos que desafían las técnicas y prácticas tradicionales (Constantiou y Kallinikos, 2015). A mayor desarrollo organizacional, mayor se complejiza la estructura y con ello el tratamiento de los datos.

En cuanto a tratamiento, el manejo de la información para la toma de decisiones sufrió cambios gracias a la incorporación de tecnologías de datos, así como también la problemática vinculada a la privacidad de los datos y el cumplimiento regulatorio. Esto se ve también impactado por el modelo de negocios al que está cambiando el mercado inmobiliario actual.

Un modelo de negocios es una representación abstracta de una organización. Lo que se intenta entender es cómo un determinado modelo de negocio de plataforma se consolidó en la última década, en contrapartida con el modelo de negocios lineal. ¿Qué es un modelo lineal? Se identifica por modelo lineal a la forma tradicional de hacer negocios en la cual la empresa es capaz de generar valor al cliente desde el primer momento; por ejemplo, una floristería, un supermercado, una peluquería o una inmobiliaria tradicional a la que un interesado en vender se acerca para llevar información de su propiedad y, en contrapartida, un interesado en comprar se acerca para que le cuenten qué propiedades están en venta.

Un modelo de negocio de plataforma genera propuestas de negocio al permitir interacciones entre personas, grupos y usuarios aprovechando efectos de red. Por lo general comprenden dos lados: la oferta y la demanda, los productores y los consumidores. Iniciar las interacciones entre esos dos lados es uno de los elementos cruciales para una plataforma.

Los modelos de plataforma tienen otra lógica en la generación de valor. Como indica Alex Moazed en su blog "Platform business model - Definition - What is it? – Explanation", para que los intercambios sucedan, las plataformas aprovechan y crean grandes redes escalables de usuarios y recursos a los que se puede acceder bajo demanda. Las plataformas crean comunidades y mercados con efectos de red que permiten a los usuarios interactuar y realizar

transacciones. Las plataformas exitosas facilitan los intercambios al reducir los costos de transacción y/o al permitir la innovación externalizada. Con la llegada de la tecnología conectada, estos ecosistemas permiten que las plataformas se escalen de formas que las empresas tradicionales no pueden.

Como ejemplo de un sistema inmobiliario de modelo de plataforma, se puede considerar la propuesta de ArgenProp. Este es un negocio basado en una plataforma de ventas, compra y alquiler de propiedades en la que compradores y vendedores, locatarios y locadores interactúan intercambiando propiedades de todo tipo. Este tipo de plataformas han conseguido posicionarse como una de las opciones favoritas de intercambio electrónico por la posibilidad de poder conseguir mayor dinamismo y visibilidad a la hora de contractar una transacción. Una única plataforma engloba muchas inmobiliarias ofreciendo sus propiedades disponibles, por lo que quien se encuentra interesado puede buscar en un solo lugar lo que necesita, sin necesidad de ir puerta a puerta por todas las inmobiliarias de la zona. Lo mismo para las inmobiliarias, que de esta manera pueden cubrir un área más grande que la de su zona cercana.

Algunas inmobiliarias lo utilizan como un segundo canal de contacto, ya que mantienen su local a la calle, mientras que otros simplemente son matriculados que utilizan la virtualidad como su único lugar de oferta. De este modo los interesados simplemente tienen que crear un perfil y subir sus propiedades, la plataforma se encargará de realizar todo el proceso de visibilidad y contacto, hasta el punto de ofrecer hasta cobertura de caución para los alquileres que intermedian.

Pese a que la tecnología cambia, las leyes de la economía no lo hacen. Se suele utilizar el término información de manera muy amplia. Esencialmente, cualquier cosa que pueda ser digitalizada es información. Datos es un término que se refiere a hechos, eventos, transacciones, etc., que han sido registrados. Es la entrada sin procesar de la cual se produce la información. Información se refiere a los datos que han sido procesados y comunicados de tal manera que pueden ser entendidos e interpretados por el receptor. Alguna información tiene valor de entretenimiento, otras de negocios, pero más allá de cuál sea la fuente de los datos, la gente está dispuesta a pagar por información. Muchas estrategias para los proveedores de información se

basan en el hecho de que los consumidores difieren mucho en cómo valoran los bienes de información en particular. Y, por supuesto, la información es costosa de crear y ensamblar.

La estructura de costos de un generador de información es bastante inusual. Dado que la naturaleza misma de la competencia en los mercados de información está impulsada por esta estructura de costos inusual, esto da comienzo a la descripción general de la estrategia de información.

La información es costosa de producir pero muy barata para reproducir. A través de la empresa (que es base para este análisis de gestión de datos) ArgenProp se están vendiendo en este momento 85.000 departamentos en la Ciudad Autónoma de Buenos Aires. Las visitas crecieron exponencialmente, llegando a 150 por segundo. Cada una de estas visitas generó, además de una erogación dinero para el quien publica y un ingreso para la empresa, miles de millones de datos que la compañía de plataforma supo aprovechar.

Los economistas dicen que la producción de buena información implica altos costos fijos pero bajos costos marginales. El costo de producir la primera copia del bien de información puede ser sustancial, pero el costo de producir (o reproducir) copias adicionales es insignificante. La estructura de costos de la información de ArgenProp involucra la generación de grandes volúmenes de información a muy bajo costo, que se generan en su propia plataforma y de manera muy rápida.

Lo costoso para ArgenProp es el aprovechamiento de dichos datos y su transformación a información. Lo primero a tener en cuenta es el repositorio de datos que debe mantener. Esto implica servidores en la nube o físicos para acumular grandes volúmenes de datos que se generan a muy alta velocidad. También es importante tener en cuenta la política de protección de datos, ya que la mayoría de los datos que manejan de sus vendedores y compradores es información sensible. Para esto, ArgenProp tuvo que desarrollar, en líneas generales, una estructura de seguridad de datos. Esto se refiere a medidas de protección de la privacidad digital que se aplican para evitar el acceso no autorizado a los datos, que se encuentran en bases de datos, la navegación dentro del sitio web, etc. La seguridad de datos también protege los datos de una posible corrupción y evita la ocurrencia de delitos cibernéticos.

Sumado a esto, fue necesario que ArgenProp invirtiera también en un equipo lo suficientemente preparado como para transformar esos datos en información. El equipo de analistas de datos, gracias a la interpretación de los mismos, pudo establecer estrategias dentro de la empresa, que son utilizadas por distintas áreas como marketing, el equipo comercial, recursos humanos y, también, para la toma de decisiones. Por lo tanto, debe saber recopilar datos a la vez que analizarlos de forma estadística.

Los mencionados anteriormente son costos fijos producción de información. Los costos fijos de son grandes, pero los costos variables de reproducción son pequeños. Esta estructura de costos genera importantes economías de escala: cuanto más se produce, menor es el costo promedio de producción. Pero hay más que solo economías de escala: los costos fijos y los costos variables de producir información tienen cada uno una estructura especial.

El componente dominante de los costos fijos de producir información son los costos irrecuperables, costos que no son recuperables si se detiene la producción. Los sueldos que se pagaron, no se recuperan. La tecnología en la que se invirtió, no puede venderse a terceros. Los desarrollos y el tiempo invertido en estos para generar la información, tampoco son bienes comercializables.

Convertirse en usuario de ArgenProp es muy simple: sólo se debe ingresar al sitio, presionar en la opción ‘Crea tu cuenta’ y completar unos pocos datos. Esto ya está generando información para la plataforma que es almacenada en su base de datos. Si quien se registra es una inmobiliaria con expectativas de seguir expandiéndose en el corto plazo, tiene la posibilidad de registrarse como empresa ingresando en ‘Crear una cuenta de desarrollador, propietario o inmobiliaria’. Esto abre un nuevo abanico de datos que es almacenado en la misma base de datos anterior, pero bajo otro perfil.

La empresa ArgenProp entiende como Base de datos Personales al conjunto organizado de Datos Personales ordenado en una base de datos que será de titularidad de la Empresa. Un Dato Personal es información de cualquier tipo referida a personas físicas o de existencia ideal que sean clientes de la Empresa. Dicha información consiste en nombre, domicilio, Documento

Nacional de Identidad, Identificación Tributaria, teléfono, dirección de correo electrónico y datos vinculados al pago del servicio (i.e., número de tarjeta de crédito, CBU, etc.).

Todas las operaciones y procedimientos sistemáticos, electrónicos o no, que permitan la recolección, conservación, ordenación, almacenamiento, modificación, evaluación, bloqueo y en general, el procesamiento de los Datos Personales conforman el tratamiento de esos Datos Personales. Para poder operar con la plataforma, el cliente conoce y acepta que ArgenProp y/o quien ésta designe expresamente a tal efecto podrá requerirle determinada información que puede ser considerada como "Datos Personales" en virtud de lo dispuesto por la ley 25.326 (Protección de Datos Personales) a efectos de la gestión comercial y publicidad. Asimismo, podrá enviarle correos electrónicos en relación con el contenido del Sitio Web, los servicios prestados, o sobre su cuenta y en respuesta a sus preguntas, pedidos, consultas o comentarios.

ArgenProp también le podrá enviar correos electrónicos con información sobre productos y servicios ofrecidos por la misma y/o terceros asociados comercialmente que le puedan resultar de interés, y para cumplir con la legislación aplicable, a menos que usted indique expresamente que no desea recibir dichos correos electrónicos a través de los procesos implementados a tal efecto y también podrá contratar a terceros para la prestación del servicio de almacenamiento, ordenación, modificación, evolución, bloqueo y en general el procesamiento de los Datos Personales.

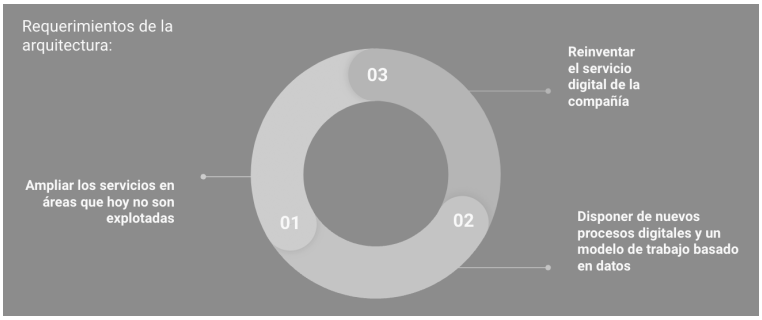
Asimismo, y en cumplimiento de la ley 25.326 y las disposiciones dictadas por la Dirección Nacional de Protección de Datos Personales, se informa al Cliente que el titular de los Datos Personales tiene la facultad de ejercer el derecho de acceso a los mismos en forma gratuita a intervalos no inferiores a seis meses, salvo que se acredite un interés legítimo al efecto conforme lo establecido en el artículo 14, inciso 3 de la ley 25.326.

Toda persona tiene derecho a que sean rectificados, actualizados y, cuando corresponda suprimidos o sometidos a confidencialidad los Datos Personales de los que sea titular, que estén incluidos en un banco de datos (artículo 16 de la ley 25.326). El titular podrá en cualquier momento solicitar el retiro o bloqueo de su nombre de los bancos de datos.

Sobre la arquitectura de datos de la organización, se propondría la aplicación de *The Open Group Architecture Framework* (TOGAF) (o Esquema de Arquitectura del *Open Group*, en español) es un esquema (o marco de trabajo) de arquitectura empresarial que proporciona un enfoque para el diseño, planificación, implementación y gobierno de una arquitectura empresarial de información. Esta arquitectura está modelada, por lo general, en cuatro niveles o dimensiones: Negocios, Tecnología (TI), Datos y Aplicaciones. Cuenta con un conjunto de arquitecturas base que buscan facilitarle al equipo de arquitectos cómo definir el estado actual y futuro de la arquitectura.

Un esquema de arquitectura es un conjunto de herramientas que puede ser utilizado para desarrollar un amplio espectro de diversas arquitecturas. Este esquema debe describir una metodología para la definición de un sistema de información en términos de un conjunto de bloques constitutivos que encajen entre sí adecuadamente, contener un conjunto de herramientas para el diseño de sistemas, usando lenguajes de modelación para definir los componentes del sistema mediante diversos tipos de diagramas, proveer un vocabulario común basado en el enfoque de sistemas (teoría genera l de los sistemas) e incluir una lista de estándares recomendados. TOGAF cumple estos requisitos. Esto sería para la implementación a futuro.

Ilustración 1: Requerimientos de la arquitectura según TOGAF



Fuente: Elaboración propia

Base de datos

Procedencia de la Base de Datos

La base de datos utilizada se descargó de la página *web* de Properati[3]. La información obtenida fue descargada el 1 de abril de 2022, que es cuando se inició el desarrollo y análisis de este trabajo. La información es de libre acceso por lo que no se infringió ninguna Ley de *Copyright* ni violación de privacidad de los datos al obtener la misma.

Limpieza y preparación de los datos

Con la utilización de *Microsoft Excel*, se concatenan los atributos *title, description, services, additionals*, bajo una sola con nombre *description*, para así tener todo lo que corresponde a texto en un único campo. La intención de esto es poder aplicar *text mining* para encontrar atributos claves en las propiedades en venta. Se eliminan entonces los atributos de texto que se concatenaron en un único campo para así también reducir el tamaño de la base de datos y que sea más eficiente la aplicación del modelo más adelante.

Continuando con la normalización del texto contenido en los datos, se dejan afuera caracteres especiales como "ñ" y las tildes para que sea más funcional la aplicación de *text mining.* Adicionalmente se reemplaza nombres coloquiales. Como ejemplo se puede mencionar "Abasto" que se reemplaza por el nombre propio correcto del barrio (Balvanera, en este caso).

Teniendo en cuenta que existe una baja cantidad de registros en moneda Pesos Argentinos, los mismos fueron eliminados de la base de datos, para no afectar las predicciones mediante la incorporación de distintos tipos de cambio y sus consiguientes transformaciones. Asimismo, también se procede a eliminar registros que tienen su latitud y longitud duplicada, por lo que quedan entonces solo 45.485. Esto hará que la base de datos sea más pequeña, pese menos y que a su vez, el modelo pueda funcionar más rápido al momento de su aplicación.

Para contar con una variable numérica por barrio, en lugar de el nombre del barrio, se reemplaza el texto por el valor del metro cuadrado promedio por barrio. Esta información se encuentra en el Apéndice I y fue obtenida de la página de Mudafy[4] el mismo día que se descargó la base de datos de Properati, para que no existan sesgos por temporalidad.

[3] https://www.properati.com.ar/data/

[4] https://mudafy.com.ar/d/valor-metro-cuadrado-en-caba-por-barrio

Terminado esto, se pasa a trabajar en un notebook en *Google Colab* utilizando lenguaje *Python* para la limpieza de los datos y la aplicación de *text mining*. El cuaderno *Colab* se encuentra a disposición en el Anexo I. Adicionalmente a la realización de *text mining*, se buscan *outliers* y se completan los datos faltantes. Respecto a la aplicación de *text mining,* lo que se busca es obtener ciertas características de la descripción de las propiedades que la harían más valiosa al momento de determinar el precio. Se está buscando entonces ponderadores de las propiedades, como ser ubicación, medios de transporte cercanos, características propias de la construcción, etc.

Sobre el análisis de *outliers*, se eliminan los registros que cumplen con determinadas condiciones que llevarían al registro a una predicción errada. Para esto se considera que no hay propiedades con más de 20 habitaciones ni más de 20 ambientes, que no existen propiedades con más de 9 baños, que la superficie total no puede ser superior a 800 metros cuadrados ni que la cubierta puede ser superior a 700 metros cuadrados y en cuanto al precio, cualquier propiedad por encima de USD 3.000.000 fue eliminada también. Los *Outliers,* también denominados "datos atípicos" que se refiere a una observación que parece ser incompatible con el resto de los datos relativos a un modelo asumido. Esto puede producir errores en la predicción, por eso es que se retiraron de la información para evitar errores.

La mayoría de los algoritmos de aprendizaje por máquina requieren valores de entrada numéricos, y un valor para cada fila y columna en un conjunto de datos. Por lo tanto, los valores perdidos pueden causar problemas para los algoritmos de aprendizaje automático.

Es común identificar los valores faltantes en un conjunto de datos y reemplazarlos por un valor numérico. Esto se llama imputación de datos, o imputación de datos faltantes. Como indican Max Kuhn y Kjell Johnson (Applied Predictive Modeling , 2013) “una técnica popular para la imputación es un modelo de vecindad *K-nearest*. Una nueva muestra se imputa encontrando las muestras en el conjunto de entrenamiento «más cercano» a ella y promedia estos puntos cercanos para completar el valor.” Se utilizó el método de imputación de *KNN* con 5 vecinos también en *Python* para que puedan aplicarse los modelos seleccionados.

Una vez finalizado el proceso en *Python,* utilizando nuevamente *Microsoft Excel* se procede a eliminar de la base de datos atributos que no aportan valor sino que generan más peso en el documento. Po este motivo se elimina la columna *ad_type* y los atributos *L1* y *L2*, así como los registros que no corresponden a CABA y los atributos *Currency* y *Price Period*. Se elimina también el atributo *description* dado que ya se utilizó para aplicar *text mining.* Todas las fechas existentes en la documentación se formatean con formato de fecha corta.

Se agrupan las 42 nuevas variables creadas mediante la utilización de *text mining* para que sean sólo 9. Esta agrupación se realiza en base a experiencia personal en la búsqueda de propiedades, valorando a qué factores son los que se pueden agrupar y que reciben mayor ponderación e interés por parte de los compradores.

Al finalizar este punto, la base de datos se encuentra en condiciones de ser utilizada para el entrenamiento de un modelo de aprendizaje automático.

Estadística descriptiva

La base de datos cuenta con 24 variables y un total de 45.270 observaciones después de efectuada la limpieza y normalización de los datos. De las 24 variables finales, 15 son numéricas mientras que las restantes 9 son variables categóricas.

Siendo que se efectuó un análisis de celdas vacías y las mismas fueron completadas con los valores faltantes, es entendible que al realizar un análisis estadístico del *dataset*, no se encontraran celdas vacías.

Tabla 3: Resumen de estadística descriptiva

Dataset statistics

Number of variables	24
Number of observations	45270
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	8.3 MiB
Average record size in memory	192.0 B

Variable types

Numeric	15
Categorical	9

Fuente: Elaboración propia en Python

Tanto la cantidad de ambientes como la de habitaciones fueron limitados durante la limpieza de los datos a 20, siendo la media de ambientes 2.77 (que se redondea a 3) y la media de dormitorios 1.97 (que se redondea a 2). El 1.2% de las propiedades que cuentan con "0" en la variable dormitorios se entiende que corresponden a oficinas o propiedades no habitables, como terrenos.

Tablas 4 y 5: Resumen de estadística descriptiva de las variables *rooms* y *bedrooms*

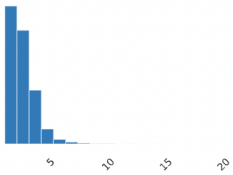
rooms

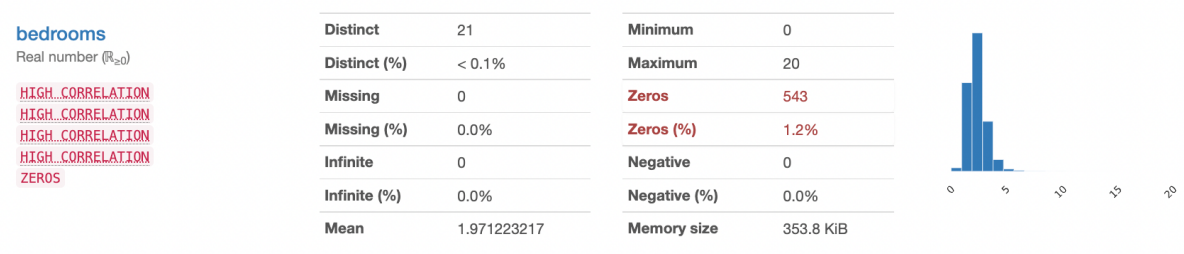
Real number ($\mathbb{R}_{\geq 0}$)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	18
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	2.775878065

Minimum	1
Maximum	20
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	353.8 KiB

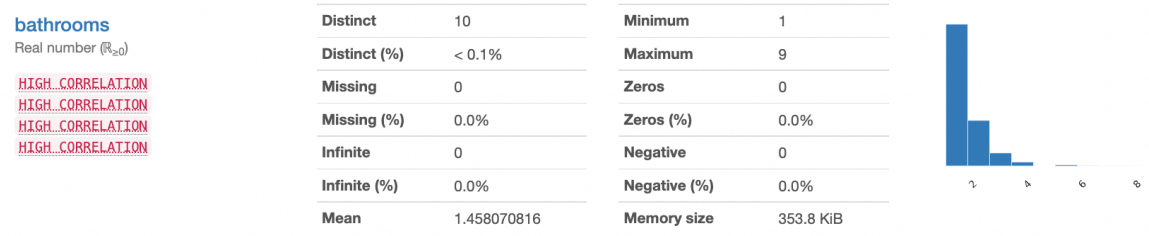




Fuente: Elaboración propia en Python

Respecto a los baños, el máximo fue limitado durante la limpieza de los datos a 9, siendo la media de baños 1.46 (que se redondea a 1 o se entiende que tiene un baño completo más un toilette, por ejemplo, pero que no son propiedades que tengan dos baños completos).

Tabla 6: Resumen de estadística descriptiva de la variable *bathrooms*

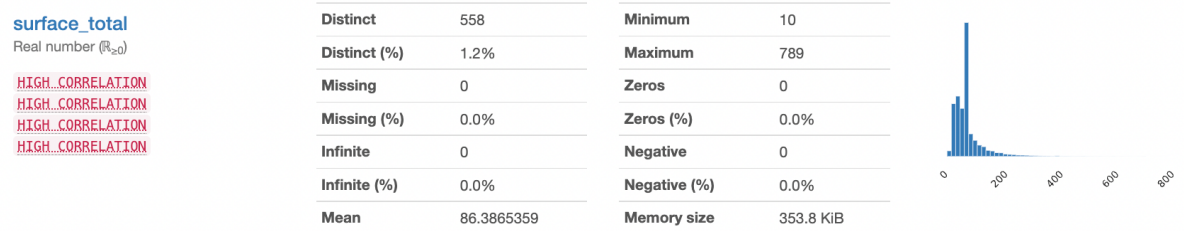


Fuente: Elaboración propia en Python

La superficie promedio de las propiedades de la base de datos analizada es de 86 metros cuadrados, mientras que la diferencia entre la mínima y la máxima es de 779 metros cuadrados. La superficie total más frecuente se encuentra entre los 35 y los 60 metros cuadrados.

En el caso de los metros cubiertos se encuentran valores similares, con una media de 76 metros cuadrados y los valores más frecuentes también entre 35 y 60 metros cuadrados.

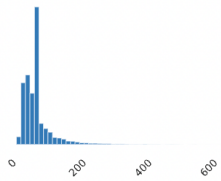
Tablas 7 y 8: Resumen de estadística descriptiva de la variable *Surface_total* y *Surface_covered*



surface_covered
Real number (ℝ≥0)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

Distinct	485	Minimum	12
Distinct (%)	1.1%	Maximum	700
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	75.52081997	Memory size	353.8 KiB



Fuente: Elaboración propia en Python

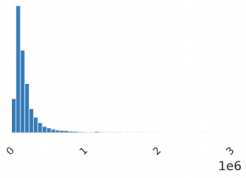
La variable precio tiene una media de USD 215.787. La misma se encuentra muy correlacionada con las variables ambientes, dormitorios, baños y cantidad de metros, tanto cubiertos como totales.

Tabla 9: Resumen de estadística descriptiva de la variable *price*

price
Real number (ℝ≥0)

HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION
HIGH CORRELATION

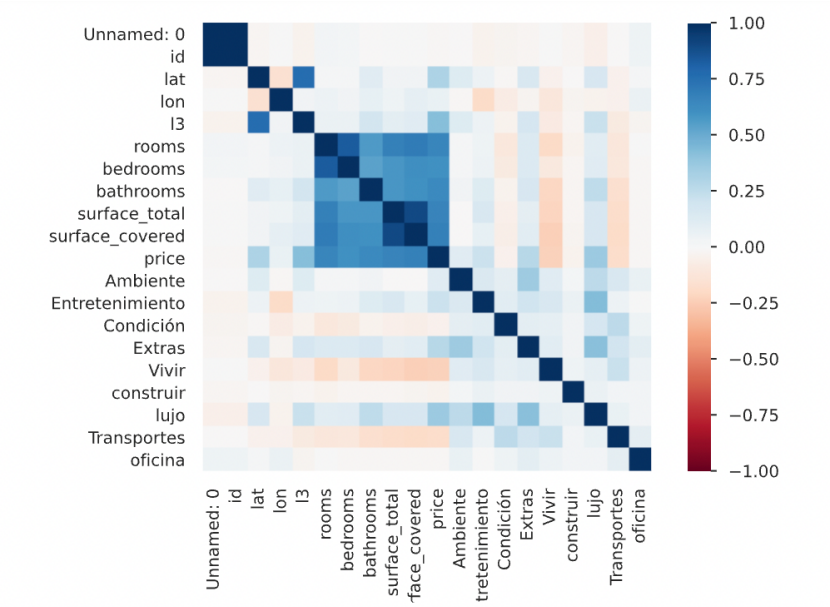
Distinct	2416	Minimum	15000
Distinct (%)	5.3%	Maximum	3000000
Missing	0	Zeros	0
Missing (%)	0.0%	Zeros (%)	0.0%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	215787.9002	Memory size	353.8 KiB



Fuente: Elaboración propia en Python

El coeficiente de correlación de Spearman es una medida monotónica de correlación entre dos variables y, por lo tanto es mejor capturando relaciones monotónicas no lineales que el coeficiente de correlación de Pearson. Su valor se encuentra entre -1 y +1, -1 indicando correlación monotónica total negativa, 0 indicando que no existe correlación monotónica y 1 indicando correlación monotónica total positiva.

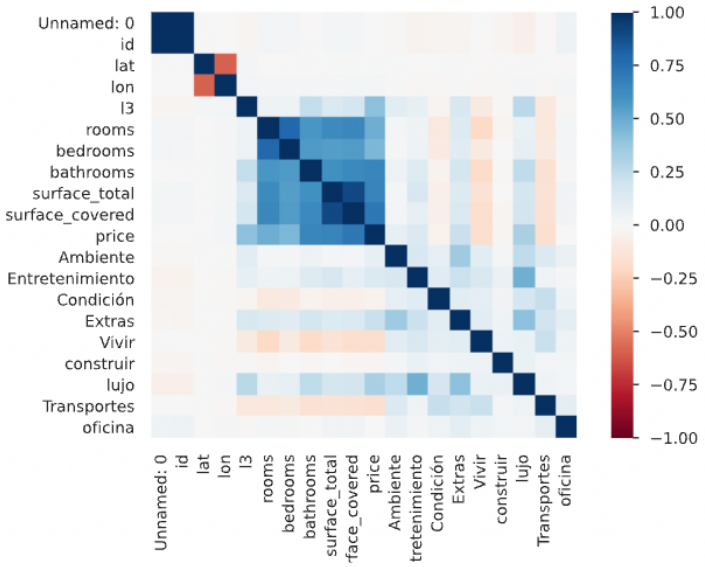
Gráfico 1: Matriz de correlación de Spearman



Fuente: Elaboración propia en Python

Lo mismo ocurre con el coeficiente de correlación de Pearson. Este es invariante bajo cambios separados en ubicación y escala de las variables, implicando que para la función lineal el ángulo a las abscisas no afecta el coeficiente.

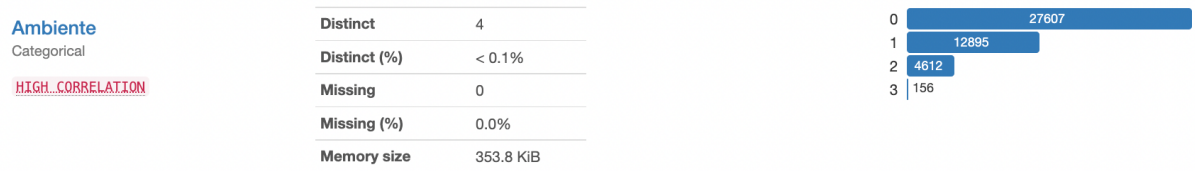
Gráfico 2: Matriz de correlación de Pearson



Fuente: Elaboración propia en Python

Sobre las variables generadas a través de *text mining,* las mismas se desarrollaron con la intención de ponderar aquellas características positivas que harían valorar más una propiedad sobre otra.

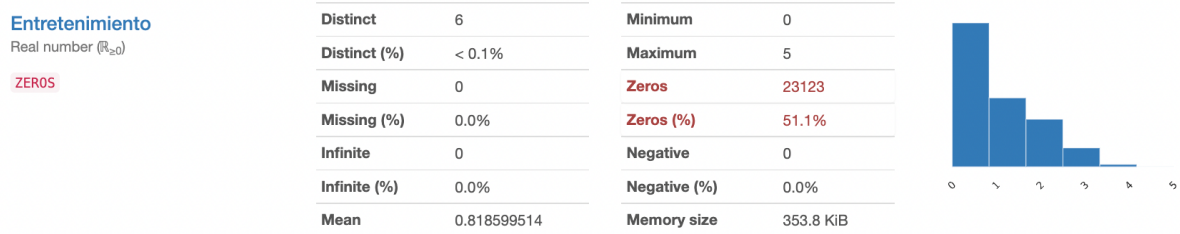
Tabla 10: Resumen de estadística descriptiva de la variable Ambiente



Fuente: Elaboración propia en Python

Ambiente, por ejemplo, es la sumatoria de aire acondicionado, calefacción y loza radiante. Si cuenta con un 3 en Ambiente, se entiende que cuenta con estas 3 características.

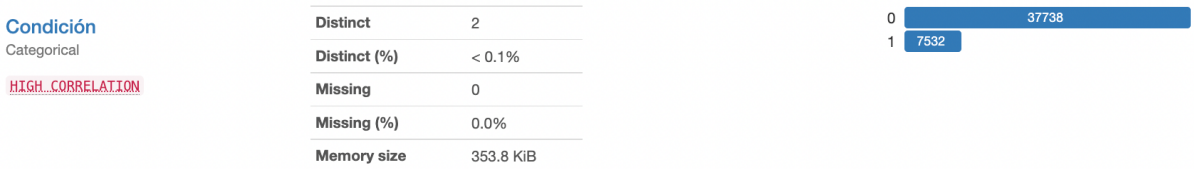
Tabla 11: Resumen de estadística descriptiva de la variable Entretenimiento



Fuente: Elaboración propia en Python

Entretenimiento es la combinatoria de parrilla, pileta, terraza, patio y piscina.

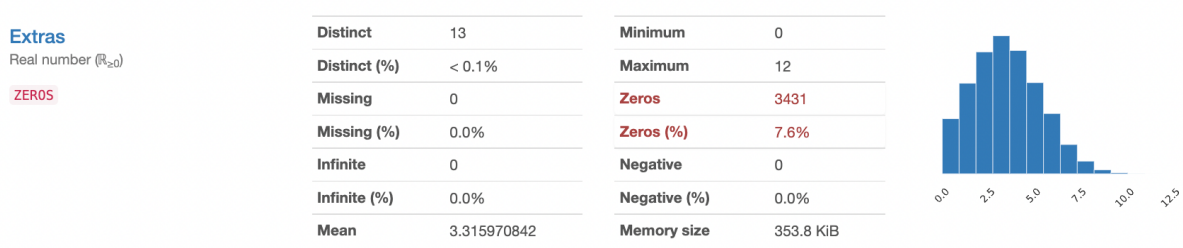
Tabla 12: Resumen de estadística descriptiva de la variable Condición



Fuente: Elaboración propia en Python

Aquellas propiedades con un 1 en condición, se entiende que son propiedades a estrenar.

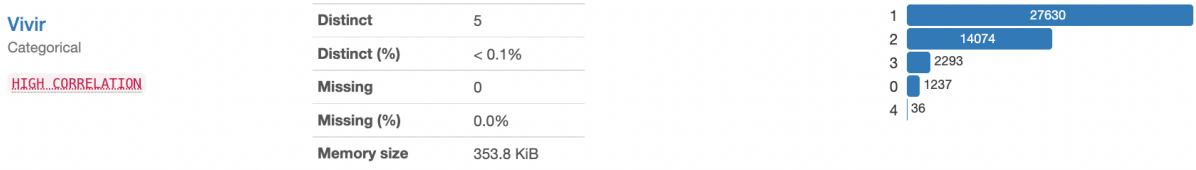
Tabla 13: Resumen de estadística descriptiva de la variable Extras



Fuente: Elaboración propia en Python

Extras es la combinación de cochera, apto mascotas, balcón, luminosidad, apto profesional, ascensor, lavadero, escritorio, vigilancia, amplitud de ambientes, vistas abiertas y ubicación en zona comercial.

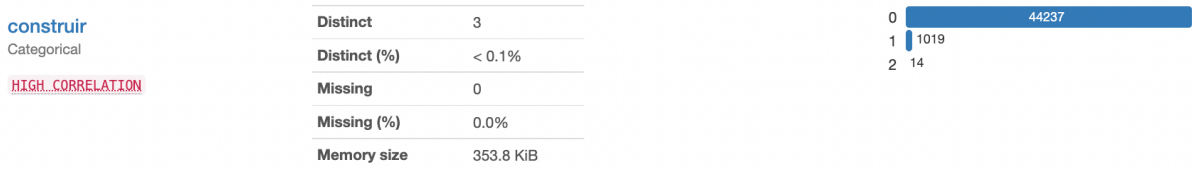
Tabla 14: Resumen de estadística descriptiva de la variable Vivir



Fuente: Elaboración propia en Python

La intención de la variable Vivir es definir aquellas propiedades que son habitables, entiéndase por eso que aquellas que tienen un 0 son oficinas.

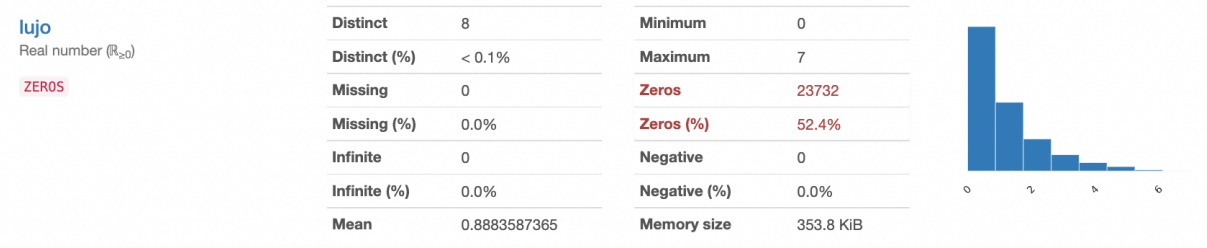
Tabla 15: Resumen de estadística descriptiva de la variable Construir



Fuente: Elaboración propia en Python

Al contrario, con la variable construir, lo que se buscó fue definir aquellas propiedades en las que es necesario construir para poder habitar, léase los terrenos o las propiedades de venta en pozo.

Tabla 16: Resumen de estadística descriptiva de la variable lujo



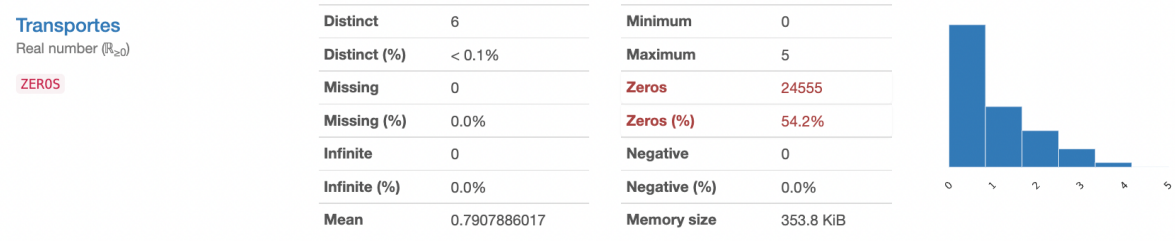
lujo
Real number ($\mathbb{R}_{\geq 0}$)
ZEROS

Distinct	8	Minimum	0
Distinct (%)	< 0.1%	Maximum	7
Missing	0	Zeros	23732
Missing (%)	0.0%	Zeros (%)	52.4%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.8883587365	Memory size	353.8 KiB

Fuente: Elaboración propia en Python

Para ponderar aquellas propiedades que tienen características de lujo, se definieron como tal con el uso de *text mining* a *Jacuzzi*, los *ammenities* que tiene una propiedad, si posee toilette, SUM, vestidor, gimnasio y pisos de porcelanato.

Tabla 17: Resumen de estadística descriptiva de la variable Transportes



Transportes
Real number ($\mathbb{R}_{\geq 0}$)
ZEROS

Distinct	6	Minimum	0
Distinct (%)	< 0.1%	Maximum	5
Missing	0	Zeros	24555
Missing (%)	0.0%	Zeros (%)	54.2%
Infinite	0	Negative	0
Infinite (%)	0.0%	Negative (%)	0.0%
Mean	0.7907886017	Memory size	353.8 KiB

Fuente: Elaboración propia en Python

Buscando dar más peso a aquellas propiedades que se encuentran bien comunicadas, las que estaban cerca de un subte, colectivo, tren, autopista u otro medio de transporte también fueron ponderadas.

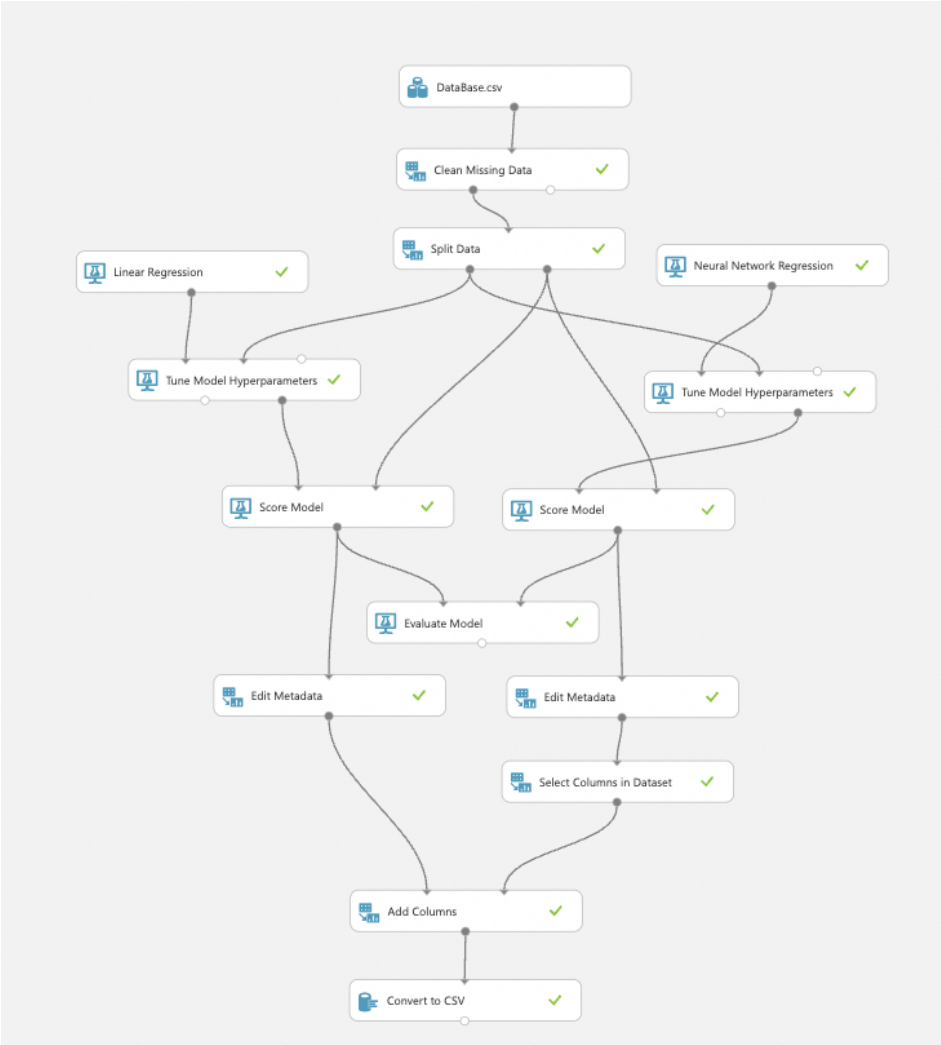
Aplicación de métodos analíticos predictivos

El presente análisis se realizó combinando herramientas como *Microsoft Excel* y *Google Colab* (lenguaje *Python*) para la preparación de la información antes de la aplicación de los modelos, como fue desarrollado en el apartado anterior. Los modelos fueron aplicados utilizando *Microsoft Machine Learning Studio (classic).*

Desarrollo y aplicación de modelos de predicción

MS Machine Learning Studio (classic)

Gráfico 3: Desarrollo de modelo de aprendizaje automático para la predicción de precios de propiedades en la Ciudad Autónoma de Buenos Aires – 2Q 2022



Fuente: Elaboración propia en Microsoft Machine Learning Studio

La *performance* de los modelos se evaluó analizando qué tanto distan las predicciones del modelo de los valores de precios reales al momento de estimar los precios de propiedades. Se aplicaron dos modelos de regresión distintos. El primero es un modelo de Regresión Lineal y el segundo es un modelo de Regresión de Red neuronal.

Con un modelo de regresión, se predice o se estima el valor numérico desconocido, de acuerdo con unas características dadas. La diferencia entre la predicción y el valor real es el error. Se compara para ambos modelos el RMSE (Error medio cuadrático) y el MAE (Error medio absoluto). Ambas métricas se utilizan para la comparación de modelos. Los datos solos no dan información.

La métrica más comúnmente utilizada para las tareas de regresión es el error medio cuadrático y representa a la raíz cuadrada de la distancia cuadrada promedio entre el valor real y el valor pronosticado. Indica el ajuste absoluto del modelo a los datos, cuán cerca están los puntos de datos observados de los valores predichos del modelo. El error cuadrático medio o RMSE es una medida absoluta de ajuste. Como la raíz cuadrada de una varianza, RMSE se puede interpretar como la desviación estándar de la varianza inexplicada, y tiene la propiedad útil de estar en las mismas unidades que la variable de respuesta. Los valores más bajos de RMSE indican un mejor ajuste. RMSE es una buena medida de la precisión con que el modelo predice la respuesta, y es el criterio más importante para ajustar si el propósito principal del modelo es la predicción.

Tabla 18: Evaluación de *performance* de resultados del modelo de regresión lineal

Metrics	
Mean Absolute Error	123936.890662
Root Mean Squared Error	274253.114048
Relative Absolute Error	0.875064
Relative Squared Error	1.11739
Coefficient of Determination	-0.11739

Metrics	
Mean Absolute Error	899281.503774
Root Mean Squared Error	916345.916716
Relative Absolute Error	6.34943
Relative Squared Error	12.474426
Coefficient of Determination	-11.474426

Fuente: Elaboración propia en Microsoft Machine Learning Studio

Esta métrica es sensible a *outliers* porque está elevada al cuadrado. la consecuencia de esto es que pondera más las diferencias. Penaliza más a un modelo por cometer errores grandes que por cometer errores chicos. Lo que es necesario analizar es ¿qué tan grave es para el mercado

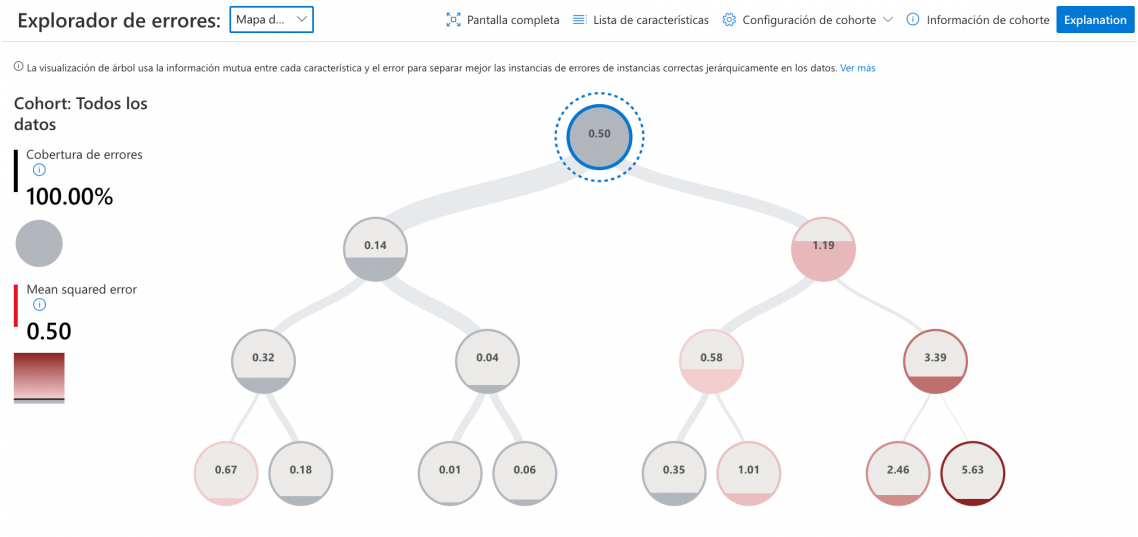
inmobiliario esta diferencia? La respuesta a esta pregunta es netamente práctica. Con un precio mal estimado se puede perder a un vendedor, porque se le está ofreciendo un precio inferior al real y puede no querer vender a ese valor o se puede estimar un precio más alto, por lo que la propiedad estará mucho tiempo a la venta generando gastos innecesarios para el propietario.

Considerando el valor de RMSE, ambos modelos tienen un coeficiente de determinación negativo. Esto significa que ninguno de los dos modelos es es aceptable. Ambas predicciones son peores que estimar la media de la base de datos.

El MAE se calcula como el promedio de la diferencia absoluta entre el valor observado y los valores predichos. El error medio absoluto o MAE es un puntaje lineal, lo que significa que todas las diferencias individuales se ponderan por igual en el promedio. Con esta medida de error se obtiene el mismo resultado, siendo que ninguno de los dos modelos predice mejor que la media.

Realizando un análisis de Error para el modelo de Regresión Lineal en *Python* utilizando el paquete de *error-analysis* se obtiene que existe error en el 100% de las estimaciones

Gráfico 4: Mapa de errores del modelo de Regresión Lineal



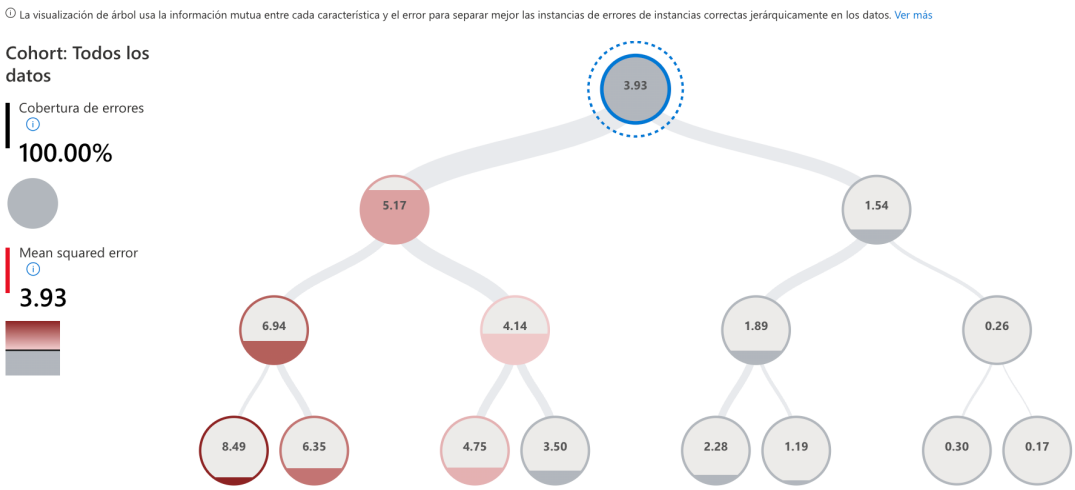
Fuente: Elaboración propia en Python

El mapa de errores muestra que el 62.93% de los errores se concentra en aquellos con precio superior a USD 205.265, mientras que el error se acentúa en aquellos de precio superior

a USD 492.500. Donde mejor desempeño tiene el modelo es en aquellas propiedades con precio entre USD 108.868 y USD 145.500. Analizando la base de datos, son propiedades de menos de 430 metros cuadrados de superficie total o 400 metros cuadrados de superficie cubierta. Son propiedades de menos de 3 ambientes en barrios más onerosos de la Ciudad Autónoma de Buenos Aires, como ser Palermo, Colegiales, Belgrano y Recoleta. Por otro lado, en barrios de valor promedio hasta USD 2.000 el metro cuadrado, predice de manera correcta para propiedades de hasta 2 ambientes. Para propiedades con una superficie mayor a 430 metros cuadrados, se recomienda solicitar a un vendedor que haga una cotización *in-situ*.

En el caso del análisis de error del modelo de Regresión de Red Neuronal, en cambio, se observa que el 78.90% de los errores se concentran en aquellas propiedades de precio superior o igual a USD 205.265. La menor cobertura de errores se da en aquellas propiedades de precio superior a USD 1.010.000. El modelo de Regresión de Red Neuronal es, por lo tanto, más exacto para propiedades de más de 400 metros cuadrados, por lo que la cotización *in-situ* realizada por un vendedor puede ser verificada con el modelo.

Gráfico 5: Mapa de errores del modelo de Regresión de Red Neuronal



Fuente: Elaboración propia en Python

Elaboración de la propuesta de valor para una inmobiliaria

Machine learning es una tecnología que se basa en estas dos fases fundamentales: la primera, la adquisición de conocimientos basados en los datos (a través del uso de *big data*); la segunda, hacer predicciones apoyadas en esos conocimientos adquiridos.

El *machine learning* funciona gracias al almacenamiento y alimentación de datos, dos actividades que han estado presentes desde hace décadas. Las computadoras son capaces de procesar y analizar rápida y eficazmente todos esos datos almacenados. Ahora es posible recopilar datos y crear algoritmos que proporcionen nuevos conocimientos, generando así un gran valor añadido.

El sector inmobiliario ha sido uno de los sectores más reacios a la digitalización, dada la tradicionalidad de este mercado. Sin embargo, la innovación se está volviendo más necesaria en este sector dadas sus posibilidades en la valoración y gestión de inmuebles.

Uno de los campos que más se están explotando en este sentido es precisamente la valoración de propiedades, basada en datos privados y públicos y en algoritmos que permiten realizar de manera objetiva ofertas transparentes y competitivas.

Mediante la entrega de dos modelos de aprendizaje automático de Regresión, cada inmobiliaria puede estimar el costo de una nueva propiedad o, para aquellos casos donde el modelo no es lo suficientemente asertivo tal como se describió en el apartado anterior, al menos asiste en verificar si el valor cotizado *in-situ* está en línea con los precios del mercado.

Conclusión

Haciendo una comparación entre el precio real de las propiedades y las predicciones obtenidas para cada uno de los modelos, se ve como los modelos de regresión aplicados no predicen de manera correcta para propiedades monoambiente.

El modelo de regresión lineal es más eficiente en departamentos de 2 y 3 ambientes, con medidas inferiores a 400 metros cuadrados, en barrios de la Ciudad Autónoma de Buenos Aires

como Belgrano, Palermo, Colegiales y Recoleta y se muestra más exacto en las propiedades de 3 o más ambientes de los barrios que tienen valor promedio de metro cuadrado inferior a USD 2.000.

El modelo de regresión de red neuronal se muestra eficiente para propiedades de 4 o más ambientes de los barrios más caros de la Ciudad, siendo también más exacto en propiedades de más de 400 metros cuadrados.

Para los casos que no se encuentran enmarcados entre los parámetros anteriores, se recomienda una cotización *in-situ.*

Los siguientes pasos para la mejora del proceso serán la prueba de otros modelos de regresión no lineal, como ser la regresión exponencial, potencial o parabólica y el análisis de su *performance* y errores para entregar modelos más eficientes a las inmobiliarias.

Referencias Bibliográficas

Lescano, R. (2019). Big Data, Machine Learning y Deep Learning: conceptos y diferencias. *Enzyme Advising Group*, https://blog.enzymeadvisinggroup.com/big-data-machine-learning.

Paul D. Berger, N. I. (1998). Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing, 12*(1), 17-30.

InmoGesco. (2022). *Big data inmobiliario: ¿Cómo usarlo en una inmobiliaria?* Obtenido de https://inmogesco.com/blog/big-data-inmobiliario/

Gamarra, V. (2020). *EL NUEVO SECTOR INMOBILIARIO (PropTech): caminando hacia el Big Data.*

Deloitte. (febrero de 2018). *Data is the new gold: the future of real estate service providers.* Obtenido de https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Public-Sector/gx-real-estate-data-new-gold.pdf

Yanfang Niu, L. Y. (2021). Organizational business intelligence and decision making using big data analytics. *Information Processing & Management*, Volume 58, Issue 6.

https://blog.powerdata.es. (diciembre de 2021). Obtenido de Blog Powerdata: https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/big-data-y-business-intelligence-motor-de-negocios-inteligentes

Power, D. J. (2007 de 2007). *A brief history of decision support systems.* Obtenido de Recuperado del sitio Web de DSSResources: http://dssresources.com/history/dsshistory.html.

Oracle. (2011). *Oracle Database VLDB and Partitioning Guide 11g Release 2 (11.2).* Obtenido de http://docs.oracle.com/cd/E11882_01/server.112/e25523.pdf

Ponniah, P. (2004). *Data warehousing fundamentals: a comprehensive guide for IT professionals.* John Wiley & Sons.
Inmon. (2005). *Building the data warehouse.* John Wiley & Sons.
Rouse, M. (2010). *Search Business Analytics.* Obtenido de http://searchbusinessanalytics.techtarget.com: http://searchbusinessanalytics.techtarget.com
SAS Institute. (s.f.). *Eight levels of analytics.* Obtenido de http://www.sas.com/news/sascom/analytics_levels.pdf
Power, D. J. (2007). A brief history of decision support systems. *DSSResources: http://dssresources.com/history/dsshistory.html, versión 4.* Obtenido de http://dssresources.com/history/dsshistory.html
Power, D. (2007). A brief history of decision support systems. *DSSResources: http://dssresources.com/history/dsshistory.html*.
Brynjolfsson, E. H. (2011). Strength in numbers: How does data-driven decisionmaking affect firm performance? *http://ssrn.com/abstract= 1819486*.
Davenport, T. H. (2009). Make better decisions. *Harvard Business Review*, 87(11), 117-123.
Davenport, T. H. (2006). Competing on analytics. *Harvard Business Review*, (84), 98-107.
Weill, P. &. (2006). Generating premium returns on your IT investments. *Sloan Management Review*, 47(2).
Johnson, M. K. (2013). *Applied Predictive Modeling* .

Anexos

Anexo I

Cuaderno Colab con la limpieza de los datos
https://drive.google.com/file/d/1ykH9x95-yCfxjVceSnjju_qgDRE4njm-/view?usp=sharing

Anexo II

Salida de Microsoft Machine Learning Studio

https://drive.google.com/file/d/11sSSdS2KwmJEWXVNDikXArcpYXVoXfQc/view?usp=sharing

Apéndices

Apéndice I - Base de datos utilizada

https://drive.google.com/file/d/1YRkphiaihjW6oSjOwR8Fj0Ca7u5ibBgq/view?usp=sharing

Apéndice II - Precio promedio por metro cuadrado – Mudafy

Barrio	Precio promedio por metro cuadrado
Agronomia	1671
Almagro	1726
Balvanera	1528
Barracas	1512
Barrio Norte	2370
Belgrano	2203
Boedo	1603
Caballito	1883
Chacarita	1722
Colegiales	1975
Congreso	1520
Constitucion	1289
Flores	1601
Floresta	1487
Florida	2042
La Boca	1209
La Lucila	2197
Liniers	1540
Martinez	2201
Mataderos	1548
Monserrat	1375
Monte Castro	1618
Nueva Pompeya	1164
Nunez	2065
Olivos	1958
Once	1345
Palermo	2121
Parque Avellaneda	1339
Parque Chacabuco	1598
Parque Chas	1683
Parque Patricios	1396
Paternal	1496
Puerto Madero	4083
Recoleta	2055
Retiro	1774
Saavedra	1977
San Cristobal	1510
San Isidro	2129
San Nicolas	1511

San Telmo	1606
Tigre	1413
Velez Sarfield	1444
Versalles	1503
Vicente Lopez	2187
Villa Crespo	1751
Villa del Parque	1740
Villa Devoto	1801
Villa General Mitre	1513
Villa Lugano	1112
Villa Luro	1584
Villa Ortuzar	1679
Villa Pueyrredon	1778
Villa Real	1635
Villa Santa Rita	1543
Villa Soldati	921
Villa Urquiza	1996

Reporte del Mentor

El presente trabajo plantea el problema de estimar el precio de venta por barrio por metro cuadrado de las futuras propiedades que entren en el mercado inmobiliario. Considero que dicho problema es muy relevante en el contexto del mercado de compra y venta de inmuebles.

El problema está identificado y definido correctamente. Se realiza una descripción adecuada de las condiciones actuales y qué se pretende resolver con este trabajo. El interrogante es consistente con la problemática planteada. La hipótesis planteada de manera preliminar se definirá a medida que profundice el estado del conocimiento acerca del tema.

El planteo del problema, los objetivos y la hipótesis se encuentran articulados con la formación de base, la carrera profesional del autor y con la especialización que está realizando. Los conceptos abordados se encuentran dentro de los contenidos de los módulos de la especialización.

Además, considero que el objetivo general propuesto de desarrollar un método predictivo que permita estimar el precio de venta por barrio por metro cuadrado es coherente con el problema planteado. Y dicho objetivo es consistente con la hipótesis de que los datos de las propiedades actualmente en el mercado permiten predecir los precios futuros de las propiedades.

Finalmente, considero que el presente trabajo alcanza el objetivo propuesto al presentar un detallado análisis de los datos y un correcto desarrollo del modelo de predicción.

Roberto Abalde