



COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI ICADE CIHS

Facultad de Ciencias Económicas y Empresariales

Eficacia y Evolución de Automated Valuation Models (AVMs): Un Enfoque de Caso Bibliográfico y Práctico

Autor: Nathan Duchene

Director: Raquel Redondo Palomo

Índice

Índice	2
Resumen	4
Palabras claves	4
Abreviaciones.....	4
Definición AVM.....	6
Tipos de AVMs.....	8
Modelos Hedónicos.....	8
Modelo de Emulación de Tasación.....	9
Index Models	10
Modelos Combinados y Modelos en Cascada	10
Historia de los AVMs.....	12
Comparando el rendimiento de diferentes AVMs	14
Desarrollar un AVM	14
Regresión de Vectores de Soporte (SVR)	16
Random Forests (RF).....	16
Extreme Gradient Boosting (XGBoost).....	17
Deep Neural Network (DNN).....	17
Estudio.....	18
Recomendaciones.....	21
Caso práctico de AVM – Modelo 1 – Regresión Lineal.....	23
Introducción.....	23
Data	23
Metodología	23
Salida y Análisis de la Regresión	24
Evaluación del Modelo	27
Limitaciones y Mejoras.....	28
Caso Práctico de AVM - Modelo 2 - Regresión Lineal	30
Introducción.....	30
Idoneidad del modelo	30
Significancia del Modelo	31

Significancia de las Variables Independientes	31
Matriz de Correlación.....	33
Evaluación del Modelo	34
Métricas Claves	34
Análisis.....	34
Standard Deviation.....	35
Variance	35
Observaciones	35
Conclusión.....	35
Comparación de Casos Prácticos de AVM - Modelo 1 y Modelo 2.....	37
Ventajas y Desventajas de los AVMs	38
Ventajas de los AVMs	38
Velocidad y Eficiencia.....	38
Rentabilidad	38
Consistencia y Objetividad	39
Perspectivas Basadas en Datos	39
Desventajas de los AVMs	40
Precisión Limitada en Propiedades Únicas	40
Calidad y Disponibilidad de Datos.....	40
Falta de análisis contextual	41
Sobredependencia en la tecnología	41
Las consideraciones éticas de los AVMs y el Big Data.....	42
Preocupaciones de privacidad	42
Transparencia.....	42
Imparcialidad y Sesgo de los AVMs	42
Rendición de cuentas	43
Impacto en la profesión de tasador.....	43
Regulación y cumplimiento normativo	44
Conclusiones.....	46
Bibliografía.....	48

Resumen

Los Modelos de Valoración Automatizada (AVMs) son herramientas basadas en software que estiman el valor de propiedades mediante algoritmos y grandes volúmenes de datos. Utilizados en bienes raíces para tareas como préstamos hipotecarios y análisis de inversiones, mejoran la eficiencia y consistencia de las valoraciones en comparación con métodos tradicionales. Los AVMs se basan en datos históricos de ventas, características de las propiedades y tendencias del mercado, entre otros, para generar valores estimados. Existen varios tipos de AVMs, incluidos los Modelos Hedónicos, de Emulación de Tasación, de Índice, Combinados y en Cascada, cada uno con sus propias metodologías. Este documento explora el desarrollo y comparación del rendimiento de diferentes AVMs, como Regresión de Vectores de Soporte (SVR), Bosques Aleatorios (RF), Potenciador de Gradiente Extremo (XGBoost) y Redes Neuronales Profundas (DNN). Además, se presentan casos prácticos utilizando Regresión Lineal Múltiple (MLR) con datos del Condado de King, Washington, EE. UU. Se destacan avances tecnológicos, integración en la industria inmobiliaria y consideraciones éticas sobre privacidad de datos y sesgo en la tasación.

Palabras claves

Automated Valuation Models (AVMs), Hedonic Models, Appraisal Emulation Model, Index Models, Blended Models, Cascade Models, Regression Analysis, Machine Learning Algorithms, Multiple Linear Regression (MLR), Support Vector Regression (SVR), Random Forests (RF), eXtreme Gradient Boosting (XGBoost), Deep Neural Networks (DNN), TEGoVA (The European Group of Valuers' Association), Real Estate Appraisal.

Abreviaciones

AVM - Automated Valuation Model

SVR - Support Vector Regression

RF - Random Forests

XGBoost - eXtreme Gradient Boosting

DNN - Deep Neural Network

IVSC - International Valuation Standard Council

TEGoVA - The European Group of Valuers' Association

IVS - International Valuation Standard

REIT - Real Estate Investment Trust

HELOC - Home Equity Lines of Credit

MLR - Multiple Linear Regression

ANN - Artificial Neural Network

R² - Coefficient of Determination

MAPE - Mean Absolute Percentage Error

nRMSE - normalized Root Mean Square Error

IAAO – International Association of Assessing Officers

Definición AVM

No existe una definición consensuada para un AVM, aunque, en términos simples, un AVM es un modelo basado en software que se utiliza, en parte, en tasaciones de bienes raíces para determinar el valor teórico de una propiedad.

“An “automated valuation model” means any computerized model used by mortgage originators and secondary market issuers to determine the collateral worth of a mortgage secured by a consumer’s principal dwelling.” (Rep. Frank, 2010)

A diferencia de las tasaciones tradicionales que dependen de inspecciones físicas y el juicio subjetivo de los tasadores humanos, los AVMs emplean algoritmos informatizados para analizar grandes cantidades de datos y generar valoraciones de propiedades. Los AVMs operan recolectando y procesando diversos puntos de datos, incluidos, entre otros: datos históricos de ventas, características de las propiedades, tendencias del mercado, registros públicos y datos geográficos. Al aplicar modelos matemáticos y técnicas estadísticas a estos datos, los AVMs producen un valor estimado para una propiedad. Estos modelos a menudo incorporan análisis de regresión, algoritmos de aprendizaje automático y otros métodos estadísticos avanzados para mejorar la precisión. El resultado se presenta típicamente con intervalos de confianza, que indican el rango dentro del cual es probable que caiga el valor de mercado real. Los AVMs son ampliamente utilizados por prestamistas, profesionales de bienes raíces y consumidores para diversos fines, incluyendo préstamos hipotecarios, evaluaciones de impuestos sobre la propiedad, análisis de inversiones y monitoreo de tendencias del mercado. Los AVMs se han convertido en una herramienta integral en la industria inmobiliaria, complementando los métodos tradicionales de tasación y mejorando la eficiencia de los procesos de valoración de propiedades.

Distinguimos dos tipos de AVMs, el primero es "AVMs Orientados al Consumidor", estos son AVMs disponibles para el público, presentes en la mayoría de los grandes sitios web de tasación en línea, un ejemplo de esto es Zillow que ofrece su modelo de valoración en línea "Zestimate" que permite a los usuarios tasar su propiedad en línea de forma gratuita. El segundo tipo de AVM es de "Grado Profesional", este modelo es utilizado por profesionales e instituciones relacionadas con la industria, un ejemplo de esto sería el uso de dichos modelos por corredores de hipotecas y para uso gubernamental. Ambos tipos de AVMs son probados a gran escala para asegurar su precisión y confiabilidad.

Los AVMs dependen de los datos; esto significa que la precisión y confiabilidad de un AVM depende en gran medida de la cantidad de datos que el modelo ha recibido y es capaz de procesar. La escasez de datos de cualquiera de los parámetros de información del modelo puede llevar a un rendimiento limitado del modelo, por lo cual veremos que los AVMs tienden a desempeñarse peor en valoraciones en áreas rurales, ya que los datos tienden a ser más escasos en dichas regiones geográficas.

La precisión de los AVMs se determina mediante diferentes métodos, pero principalmente a través de pruebas comparativas a gran escala, en este caso se harán estimaciones a través del modelo y se compararán con el valor real de un precio de venta en la vida real. Las comparaciones a gran escala ayudan a determinar la precisión del modelo y a establecer puntajes de confianza. Dichos puntajes de confianza son estimaciones estadísticas de la precisión del modelo, a menudo, pero no exclusivamente basadas en el rendimiento anterior del modelo. Los puntajes de confianza establecen una referencia sobre la precisión esperada de la valoración, los puntajes de confianza a menudo se verán afectados por, pero no exclusivamente: la precisión anterior del modelo, la disponibilidad de datos y el grado de relevancia de los datos disponibles. Cuanto mayor sea el puntaje de confianza del modelo, mayor será la probabilidad estadística de que la valoración sea precisa.

En resumen, la IAAO define los Modelos de Valoración Automatizada de la siguiente manera:

“A mathematically based computer software program that market analysts use to produce an estimate of market value based on market analysis of location, market conditions, and real estate characteristics from information that was previously and separately collected. The distinguishing feature of an AVM is that it is a market appraisal produced through mathematical modeling. Credibility of an AVM is dependent on the data used and the skills of the modeler producing the AVM. AVMs should be developed by appropriately qualified market analysts, e.g. appraisers/ valuers, who use statistically-based applications to analyze data and select the best simulation of market activity for the analysis of location, market conditions, and property characteristics from previously collected data. AVMs are designed to generate value estimates for properties at specified points in time (retrospective or prospective dates as required by client).”(Dettbarn, Davis, L. Escobar, Finnsson, Ripperger, Clark, 2003).

Tipos de AVMs

La Asociación de Banqueros Hipotecarios distingue cinco tipos diferentes de AVMs (*The State of Automated Valuation Models in the Age of Big Data*, n.d.) de la siguiente manera: Modelos Hedónicos, Modelo de Emulación de Tasación, Modelos de Índice, Modelos Combinados y Modelos en Cascada.

Modelos Hedónicos

La Alianza Europea de AVM define los AVMs de la siguiente manera:

Los Modelos Hedónicos son ecuaciones matemáticas con parámetros pre-calculados que aceptan múltiples características de la propiedad como entradas (por ejemplo, tipo de propiedad, superficie, año de construcción, número de dormitorios) para calcular el valor de una propiedad. Los Modelos Hedónicos son métodos multivariados ya que normalmente utilizan múltiples parámetros para estimar los valores de las propiedades. A veces, los Modelos Hedónicos también se denominan 'Modelos de Valoración Automatizada Hedónicos' (Hedonic AVM), aunque no necesariamente llevan a cabo ningún proceso automatizado. (Statistical Valuation Standards - European AVM Alliance, n.d.)

Los AVMs Hedónicos se utilizan en tasaciones inmobiliarias para estimar el valor de una propiedad. Son específicos en el método de fijación de precios, ya que el término "hedónico" se refiere a derivar valor de las características y atributos individuales de la propiedad. Los modelos hedónicos asumen que el valor de una propiedad puede descomponerse en el valor de los atributos individuales de la misma, incluyendo, pero no limitándose a la ubicación, tamaño, número de habitaciones y fecha de construcción. Cada atributo contribuye individualmente al valor total de la propiedad. La contribución de cada atributo se estima mediante análisis estadístico de los datos del mercado.

Los AVMs Hedónicos se basan en gran medida en el Análisis de Regresión, donde los modelos de regresión estiman la relación entre el atributo y los precios de mercado. A cada atributo se le asigna un valor y contribuye a la estimación del valor total de la propiedad. Este modelo también permite ajustes que incluyen, pero no se limitan a, la inflación, ciclos económicos y variaciones del mercado inmobiliario. Los modelos hedónicos proporcionarán "márgenes de confianza", un rango de estimación estadística en el cual es más probable que el valor de la propiedad refleje la realidad del mercado. La complejidad de los Modelos Hedónicos variará dependiendo de la complejidad del análisis de datos y del número de puntos de "características de la propiedad" utilizados en el análisis.

Los modelos hedónicos tienen la ventaja de analizar el valor de la propiedad dividiendo los diferentes factores y atributos que la componen, por lo tanto, los modelos hedónicos proporcionan una comprensión matizada de lo que impulsa el valor de la propiedad. Son positivamente flexibles ya que pueden adaptarse a diferentes mercados y tipos de propiedades, ya que pueden recalibrarse fácilmente. El análisis estadístico de grandes volúmenes de datos permite estimaciones consistentes y libres de sesgo. Sin embargo, los modelos hedónicos están en gran medida sujetos a la disponibilidad de datos, por lo que la consistencia y precisión de estos modelos dependen en gran medida de la calidad de los datos y pueden ser inexactos en casos de "atributos únicos" o en áreas donde los datos son escasos.

Las aplicaciones de los Modelos Hedónicos pueden variar, aunque se utilizan en los procesos de aprobación de préstamos hipotecarios, la determinación de impuestos sobre la propiedad, la previsión de los rendimientos esperados en propiedades de inversión y la investigación de mercado destinada a detectar tendencias del mercado y estrategias de fijación de precios.

Modelo de Emulación de Tasación

Un Modelo de Emulación de Tasación *emula el proceso de una tasación manual, pero con el uso de análisis estadístico* (The State of Automated Valuation Models in the Age of Big Data, n.d.). A diferencia de los Modelos Hedónicos, los Modelos de Emulación de Tasación buscan imitar el proceso de tasación humana de bienes raíces. A diferencia de otros AVMs, estos modelos no se basan únicamente en el análisis estadístico, sino que también incorporan datos cualitativos. Por lo tanto, estos modelos intentarán comprender y reproducir el proceso de toma de decisiones de los tasadores profesionales. Los Modelos de Emulación de Tasación están diseñados para replicar las sutilezas del juicio humano, de manera que se construyen para replicar el enfoque de un tasador en la evaluación de las características de una propiedad. Si bien este modelo incorporará datos similares a los del modelo hedónico, los Modelos de Emulación de Tasación también incluirán factores cualitativos en la estimación del valor, ejemplos de esto serían, pero no se limitarían a, la condición de la propiedad, el atractivo estético y el análisis de ventas comparables.

El análisis de ventas comparables es un factor clave de los Modelos de Emulación de Tasación. En un intento por reproducir los métodos de valoración humana, este modelo atribuye un peso significativo a las ventas de propiedades comparables en el proceso

de valoración. Los datos de las propiedades comparables se clasificarán por relevancia y similitud para determinar el valor de una propiedad.

El modelo de emulación de tasación luego utiliza un análisis de datos sofisticado, técnicas estadísticas y ponderaciones para ajustar los precios de venta de las comparables, teniendo en cuenta las diferencias en características, tamaño y las tendencias históricas de precios en el área. (The State of Automated Valuation Models in the Age of Big Data, n.d.)

La combinación de datos cualitativos y cuantitativos permite una replicación del proceso humano de valoración.

Index Models

Los Modelos de Índice AVM estiman los precios de una propiedad basándose en índices agregados del mercado inmobiliario. Estos modelos rastrean las tendencias del mercado y los movimientos de precios y los aplican a las propiedades individuales. Los modelos de índice se basan en gran medida en los movimientos promedio del mercado en áreas geográficas específicas. Por lo tanto, los índices son el factor determinante de la valoración estimada de la propiedad, utilizando tantos datos históricos como tendencias del mercado. Por ejemplo, un cierto porcentaje de cambio en los precios del mercado en un área geográfica restringida permitiría al modelo estimar el cambio en el precio de una propiedad individual dentro de esta área geográfica.

Una metodología común utilizada por los Modelos de Índice es "ventas repetidas"; en este caso, los precios de la misma propiedad vendida en diferentes momentos son los puntos de datos que reflejan la evolución del mercado y, por lo tanto, la base de datos para la valoración del modelo. Estos modelos son elogiados por su simplicidad y transparencia, ya que son formas rentables y eficientes de reflejar el mercado. Sin embargo, estos modelos están limitados por su dependencia en la Calidad del Índice, y también son lentos para ponerse al día, ya que pueden no capturar la volatilidad del mercado tan distintamente como otros modelos. En comparación con otros modelos, los modelos de índice, aunque más simples, tienden a ofrecer una valoración generalizada limitada y, por lo tanto, tienden a ser menos precisos o a proporcionar márgenes de confianza más amplios.

Modelos Combinados y Modelos en Cascada

Los Modelos Combinados incorporarán parte o todos los tres modelos mencionados anteriormente. Estos modelos son adaptados por el usuario a requisitos y usos

específicos, incorporando ciertos o todos los aspectos previamente mencionados. El usuario puede adaptar estos modelos a las restricciones y especificidades de la propiedad que desea estimar; por ejemplo, la valoración de una propiedad donde ciertos puntos de datos son escasos puede requerir una combinación de las diferentes soluciones ofrecidas por los AVMs individuales, creando así una “combinación” de modelos que permite valoraciones más precisas.

Los modelos en cascada también son una forma de combinar los modelos mencionados anteriormente. En este caso, se eligen y clasifican varios modelos para evaluar el valor de una propiedad. Cada modelo proporciona su valoración, y este modelo elige los resultados más pertinentes según sus parámetros. Por ejemplo, si el modelo con la clasificación más alta proporciona una estimación que no cumple con los parámetros de nivel de confianza, entonces el modelo en cascada elegirá el siguiente modelo de clasificación más alta con el nivel de confianza aceptable como una valoración pertinente para la propiedad.

Historia de los AVMs

Los AVMs se lanzaron por primera vez a principios-mediados de la década de 1990; estas eran versiones rudimentarias de los modelos disponibles hoy en día. Los AVMs se crearon inicialmente como herramientas para aumentar la eficiencia; eran desarrollos internos que las empresas lanzaban de manera privada con la esperanza de reducir la carga de trabajo y aumentar la precisión de las valoraciones. La falta de disponibilidad de datos informatizados hacía que estos modelos fueran poco fiables. Las empresas, aunque desarrollaban estos modelos, eran reacias a utilizarlos como único modelo de valoración, y las valoraciones proporcionadas por los modelos tenían resultados de baja confianza. Además, los primeros AVMs eran lentos, como lo describió John Wierks (Babb, n.d.) :

El modelo AVM requería que los datos de la propiedad a nivel del condado se accedieran a través de un módem telefónico, conectándose a computadoras centrales para cada búsqueda individual de comparables. Ese proceso podía tomar cinco minutos por cada búsqueda.

El factor determinante que permitió el desarrollo de los AVMs no fue la necesidad de tales modelos por parte de los tasadores, sino la necesidad de los prestamistas y corredores hipotecarios de poder determinar rápidamente el valor de las propiedades para aprobar préstamos. A principios de la década de 2000, los AVMs estaban en plena expansión y las grandes empresas comenzaron a ofrecer tales modelos a los tasadores de bienes raíces. Aunque mucho más confiables que los primeros AVMs, estas nuevas y mejoradas versiones no siempre eran precisas. Como resultado, los tasadores y corredores hipotecarios comenzaron a usar AVMs en cascada, donde utilizaban múltiples modelos AVM que "rotaban" de un modelo a otro hasta que uno devolvía una valoración con un cierto nivel de confianza. (Babb, n.d.) Esto aumentó aún más el uso de los AVMs. La expansión de la disponibilidad y calidad de los datos, así como el creciente número de puntos de datos dentro de los AVMs y el aumento de la velocidad de computación y aprendizaje automático, permitieron una aceptación mucho mayor de los AVMs como métodos de valoración confiables y de confianza.

La crisis de 2008 llevó a un cambio en el uso de los AVMs, donde se usaron por primera vez AVMs "combinados", que se adaptaron aún más a la propiedad tasada y, como tal, proporcionaron resultados más confiables. Estas soluciones "combinadas" demostraron aún más su uso en tiempos más recientes, especialmente durante la crisis de COVID-19, donde se usaron ampliamente en los EE. UU. para aprobaciones hipotecarias.

Los AVMs dependen de los datos; por lo tanto, a medida que avanzamos hacia una mayor eficiencia en el análisis de datos y los modelos pueden ingerir y procesar más y más datos, podemos esperar ver un crecimiento continuo en el uso de los AVMs. La creciente disponibilidad de datos y puntos de datos permite un aumento continuo en los puntajes de confianza de los AVMs, lo que significa que los AVMs se están volviendo cada vez más confiables y, como tal, una herramienta pertinente e importante para valorar propiedades.

Si bien la necesidad de tasadores para valorar propiedades no desaparecerá en el corto plazo, podemos esperar que, con el creciente desarrollo del procesamiento de datos y la aparición de la inteligencia artificial, veamos, en un futuro cercano, un mayor aumento en el uso de Modelos de Valoración Automatizada y su capacidad para reflejar valoraciones de ventas reales.

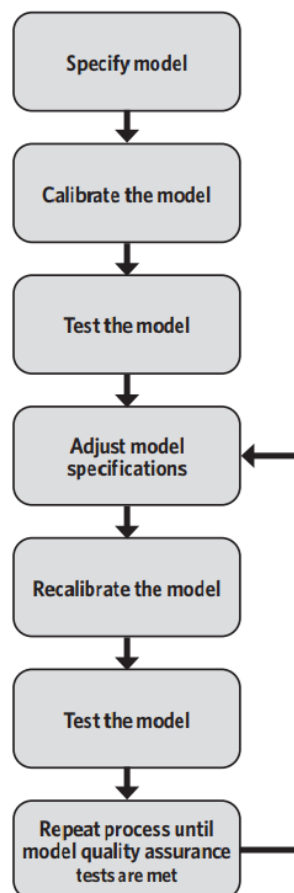
Comparando el rendimiento de diferentes AVMs

Esta sección se basa en gran medida en el artículo *Automated land valuation models: A comparative study of four machine learning and deep learning methods based on a comprehensive range of influential factors* (Jafary et al., 2024).

Diferentes AVMs producirán resultados diferentes, esto es inherente debido a las diferentes formas en que los modelos procesan los datos, el número de puntos de datos que analizan, su complejidad y su capacidad para corregir errores. Las diferencias no solo llevarán a valoraciones diferentes, sino también a diferentes rangos de confianza y precisión.

El artículo mencionado anteriormente investiga el rendimiento de cuatro AVMs diferentes: Regresión de Vectores de Soporte (SVR), Potenciador de Gradiente Extremo (XGBoost), Bosque Aleatorio (RF) y Red Neuronal Profunda (DNN). Cada AVM presenta un método de aprendizaje y procesamiento diferente, el artículo compara el rendimiento de cada uno y su capacidad para valorar con precisión los bienes inmuebles.

Desarrollar un AVM



(Dettbarn, Davis, L. Escobar, Finnsson, Ripperger, Clark, 2003)

Construir, ajustar y probar un Modelo de Valoración Automatizada no es una tarea fácil; hay varios factores a tener en cuenta, posibles complicaciones y desafíos que superar. El alcance del trabajo debe definirse, lo que significa que se deben determinar los límites geográficos del AVM (aunque teóricamente plausible, un AVM global sería altamente impráctico y posiblemente inexacto).

La base del AVM son los datos que utiliza, identificar y adquirir datos es una fase clave del desarrollo del AVM. El Estándar sobre Modelos de Valoración Automatizada distingue "tres categorías amplias: datos de propiedad, datos de ubicación y datos de mercado. Los datos de propiedad están compuestos por elementos que representan atributos físicos de la propiedad. Los datos de ubicación consideran la demografía del mercado, el tráfico, las políticas de uso del suelo y otros factores geográficos. Los datos de mercado incluyen información de ventas, ingresos y costos de reemplazo."(*Standard_on_Automated_Valuation_Models.Pdf*, n.d.). Los proveedores de datos pueden variar, aunque en la mayoría de los casos los datos se recopilan a través de: instituciones privadas, registros públicos, registros gubernamentales (disponibles para el público y para uso comercial) y datos internos (recopilados y utilizados por el desarrollador del modelo).

La estratificación es una fase clave de la "limpieza" de datos donde las propiedades se organizan en grupos homogéneos basados en factores como el uso, las características físicas o la ubicación. (*Standard_on_Automated_Valuation_Models.Pdf*, n.d.). Aunque la estratificación está normada, es subjetiva al desarrollador; la idea clave es dividir los datos en grupos manejables y coherentes para optimizar el rendimiento del AVM. La sobreestratificación o la subestratificación conducirá a variaciones limitadas en los datos, lo que obstaculiza el rendimiento del AVM.

Cada modelo está adaptado, los AVMs pueden tener un rendimiento similar utilizando diferentes procesos. Existen varias soluciones matemáticas para desarrollar un AVM, ya sea utilizando suma ponderada, no lineal, regresiones, etc. Veremos más adelante que comparar modelos determinará cuál proporciona los mejores rendimientos.

Los modelos se desarrollan a través de un ciclo de pruebas, calibración, pruebas, recalibración, etc. La idea es que el modelo gane precisión con cada ciclo, ya que se enseña y corrige a sí mismo para corregir errores y mejorar futuros rendimientos. Esto se hace a menudo mediante el aprendizaje automático, donde grandes cantidades de datos cualitativos permiten modelos de alto rendimiento (ver "Estudio").

Una vez que los modelos están suficientemente calibrados y probados, pueden aplicarse a la valoración de propiedades, proporcionando valores dentro de los niveles de confianza dados por la precisión global del modelo, la disponibilidad de datos y el rendimiento general.

En la siguiente sección, profundizaremos en el análisis comparativo de cuatro modelos diferentes. Esta comparación arrojará luz sobre las fortalezas y debilidades de cada enfoque, proporcionando información sobre las estrategias más efectivas para desarrollar Modelos de Valoración Automatizada de alta precisión.

Regresión de Vectores de Soporte (SVR)

La Máquina de Vectores de Soporte (SVM), que se basa en la teoría del aprendizaje estadístico de Vapnik, se utiliza principalmente para fines de clasificación (Vapnik, 2013). El algoritmo usado por (Jafary et al., 2024) fue modificado para procesar problemas de regresión, teniendo así la capacidad de comprender relaciones no lineales dentro de los datos (Jafary et al., 2024; Vapnik, 2013).

La Regresión de Vectores de Soporte (SVR) extiende el marco de SVM para resolver problemas de regresión. A diferencia de las técnicas de regresión tradicionales que minimizan el error directamente, SVR busca minimizar una función de costo que considera tanto la complejidad del modelo (mediante regularización) como la tolerancia al error dentro de un margen especificado. Este enfoque es particularmente útil para manejar conjuntos de datos de alta dimensión y no lineales, donde la relación entre variables es compleja y no fácilmente capturada por modelos lineales.

La robustez de SVR radica en su flexibilidad y precisión. Esta adaptabilidad hace que SVR sea una herramienta poderosa para varias tareas de regresión, incluida la valoración de bienes raíces, donde capturar las sutilezas en los datos es crucial para generar estimaciones precisas. Además, la capacidad de SVR para gestionar valores atípicos y el ruido en los datos es una ventaja significativa. Al incorporar una función de pérdida que es insensible a pequeños errores, SVR asegura que las desviaciones menores no afecten desproporcionadamente al modelo, manteniendo así la robustez y estabilidad. La aplicación de SVR en los Modelos de Valoración Automatizada resalta su potencial para proporcionar valoraciones precisas de propiedades mediante el aprovechamiento de técnicas sofisticadas de aprendizaje automático para procesar y analizar grandes cantidades de datos.

Random Forests (RF)

Este método se deriva del uso de árboles de decisión, en los cuales los conjuntos de datos se dividen secuencialmente "en subconjuntos más pequeños". El sistema opera decisión por decisión, donde en cada punto de decisión el algoritmo elige la característica y el valor de división que minimiza la varianza de la variable objetivo (Jafary et al., 2024). Este método se considera fácil de desarrollar y presenta la ventaja de la robustez, ya que el algoritmo permite mitigar problemas de sobreajuste. (Jafary et al., 2024)

Los Bosques Aleatorios se basan en el concepto de árboles de decisión creando un conjunto de múltiples árboles para mejorar el rendimiento predictivo y controlar el sobreajuste. Cada árbol en el bosque se entrena con una muestra diferente de los datos. La robustez de los RF radica en su colectividad de predicciones de los árboles individuales, lo que ayuda a suavizar la varianza y reducir el riesgo de sobreajuste que podrían encontrar los árboles de decisión individuales. Este enfoque de conjunto aprovecha el principio de la "sabiduría de la multitud", lo que resulta en un modelo que generaliza mejor a los datos no vistos. Los RF son particularmente ventajosos porque pueden manejar grandes conjuntos de datos con un mayor número de dimensiones y mantener el rendimiento sin requerir una sintonización de parámetros significativamente mayor. Además, los Bosques Aleatorios son menos sensibles al ruido en los datos, gracias a su proceso de promediado, lo que mejora aún más su estabilidad y fiabilidad. Esto los hace adecuados para aplicaciones como los Modelos de Valoración Automatizada, donde los datos pueden ser ruidosos.

Extreme Gradient Boosting (XGBoost)

El Potenciador de Gradiente Extremo (XGBoost) fue introducido por Chen y Guestrin (2016), un método que mejora el rendimiento de los Bosques Aleatorios (RF), permitiendo una mayor capacidad de generalización y un aumento general en el rendimiento del modelo RF. *XGBoost añade iterativamente árboles de decisión al modelo, optimizando la función objetivo mediante el descenso de gradiente. Controla el tamaño del paso de las actualizaciones utilizando una tasa de aprendizaje, mejorando la convergencia y el rendimiento.* (Jafary et al., 2024).

XGBoost es una implementación escalable y precisa de las máquinas de potenciación de gradiente, que mejora los métodos de potenciación tradicionales al incorporar técnicas de regularización para prevenir el sobreajuste. Este método añade secuencialmente nuevos modelos (árboles) para corregir errores cometidos por los anteriores, mejorando así la precisión de las predicciones. La naturaleza robusta y eficiente de XGBoost lo hace particularmente adecuado para conjuntos de datos complejos y tareas de aprendizaje automático a gran escala, incluyendo los AVMs. Su capacidad para manejar relaciones no lineales e interacciones entre características mejora la precisión y fiabilidad de las valoraciones de propiedades.

Deep Neural Network (DNN)

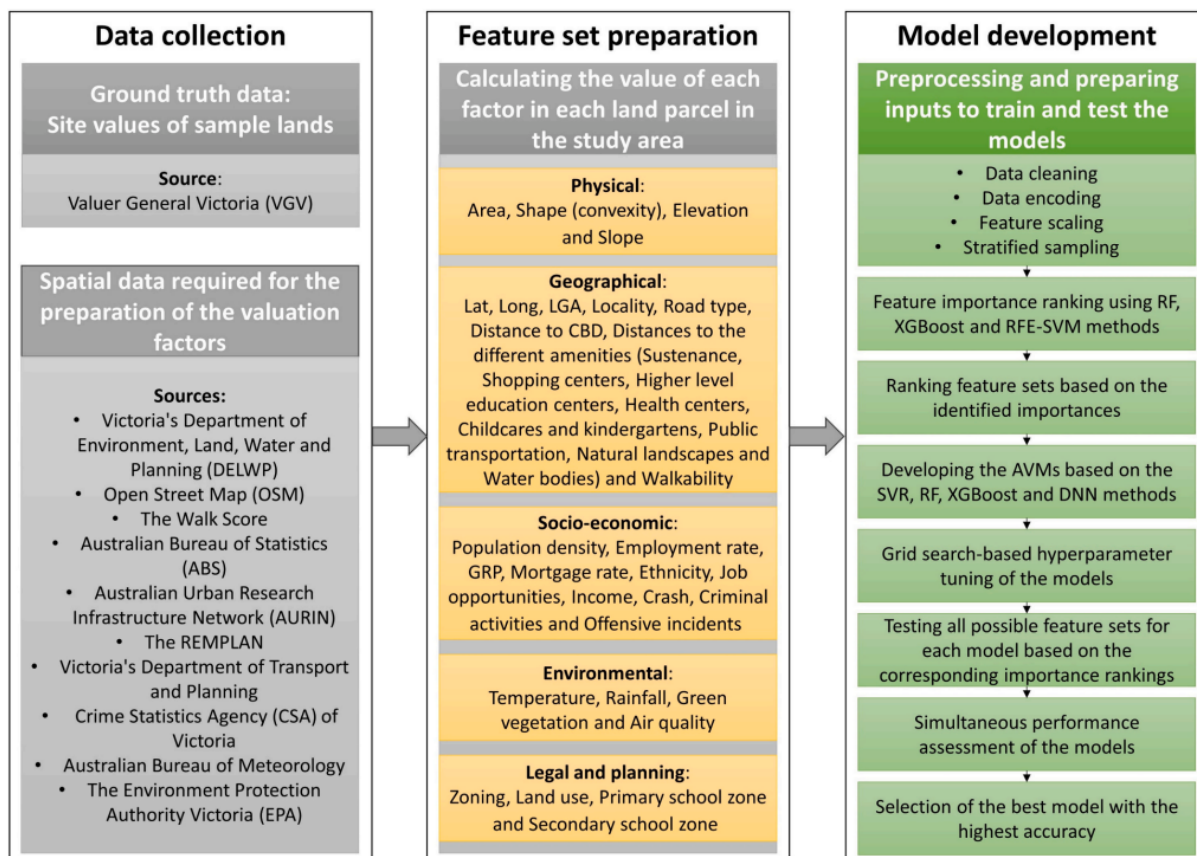
DNN es una red neuronal no tradicional que presenta la ventaja de ser capaz de aprender y procesar más datos. DNN tiene una mayor precisión que las máquinas de regresión tradicionales (ANN) y están menos sujetas al sobreajuste. DNN proporciona una solución a problemas de regresión que ANN no puede.

Las DNNs permiten capturar patrones e interacciones intrincados dentro de los datos. Las DNNs modelan funciones no lineales complejas de manera más efectiva que las redes superficiales o las técnicas de regresión tradicionales. Una DNN extrae características de alto nivel de los datos de entrada brutos, refinando progresivamente la representación y mejorando la precisión predictiva. Una ventaja clave de las DNNs es su capacidad para generalizar bien desde los datos de entrenamiento a los datos no vistos, reduciendo el riesgo de sobreajuste. Esto se logra a través de técnicas como el dropout, la normalización por lotes y la regularización (Garbin et al., 2020), que ayudan a asegurar que el modelo no se ajuste demasiado a los datos de entrenamiento. Como resultado, las DNNs pueden ofrecer un rendimiento confiable incluso con conjuntos de datos grandes y complejos. Las DNNs, combinadas con potentes GPUs de cálculo, permiten procesar grandes cantidades de datos y aprender de ellos de manera efectiva. Además, las DNNs pueden integrar diversas fuentes de datos, como características de la propiedad, factores de ubicación y tendencias del mercado, lo que ilustra la pertinencia de las DNNs para desarrollar Modelos de Valoración Automatizada.

Estudio

Los datos del estudio se recopilan a través de los archivos de las autoridades australianas "Valuer General Victoria" sobre ventas de terrenos. Su objetivo es entrenar y evaluar el rendimiento de los Modelos de Valoración Automatizada mencionados anteriormente en terrenos del departamento de Victoria, Australia.

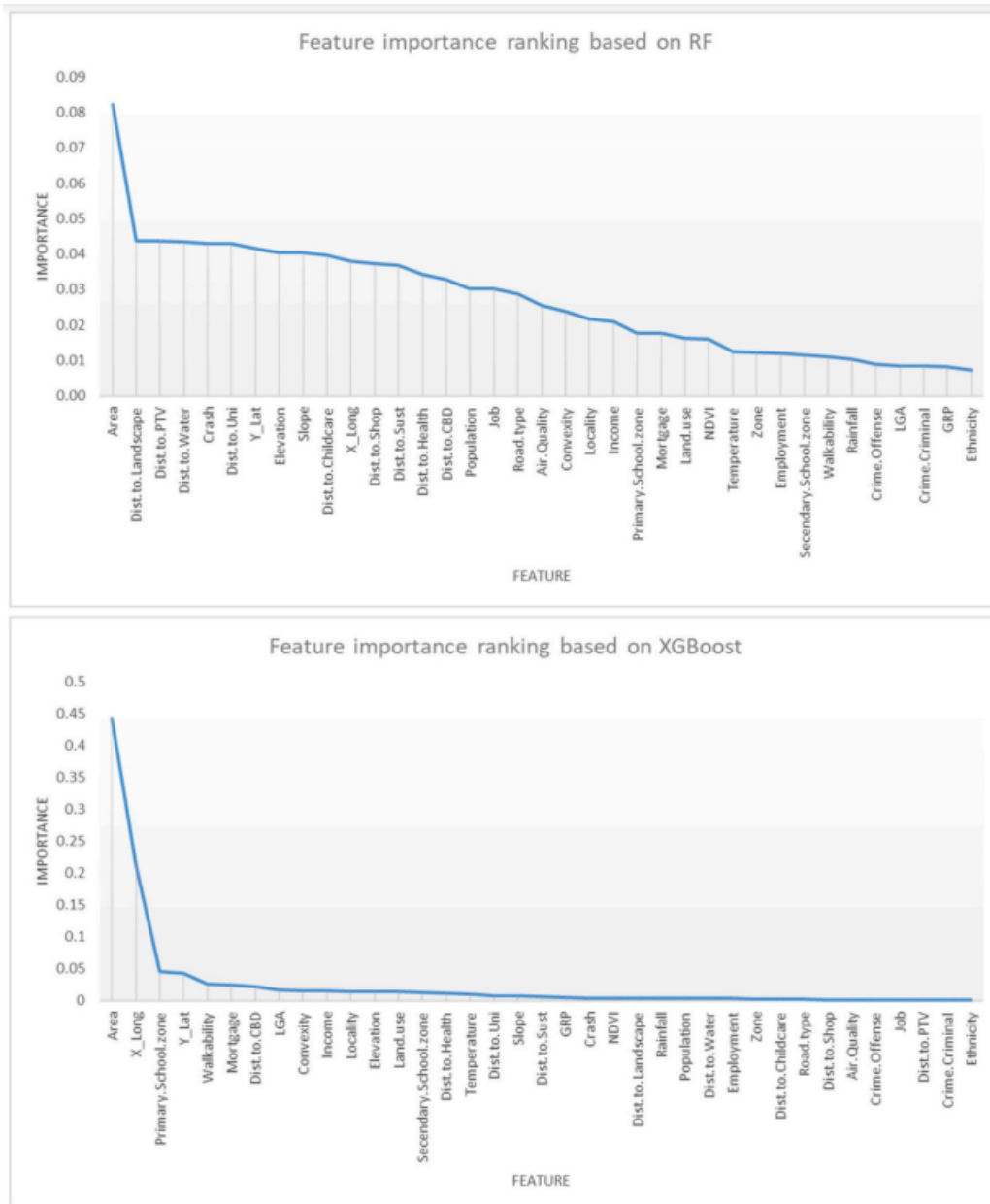
El siguiente diagrama de flujo muestra el proceso de recopilación de datos, preparación de datos y procesamiento de datos del estudio.



Flow Chart. (Jafary et al., 2024)

La segunda columna enumera los diferentes puntos de datos considerados en el estudio, un total de 37 que se ingresan en cada modelo. La tercera columna explica el proceso de desarrollo del modelo, es importante notar que los datos han sido "limpiados" y adaptados a cada modelo para asegurar el rendimiento máximo de cada uno.

Cada modelo es diferente, y por lo tanto, cada modelo atribuirá diferentes pesos a las entradas de datos. Aunque este es el caso, esto no impacta significativamente en la diferencia de la métrica de rendimiento o la valoración final, como veremos más adelante. Los siguientes gráficos muestran los pesos de las entradas por los modelos RF y XGBoost.



(Jafary et al., 2024)

Las métricas de rendimiento utilizadas en el estudio son las siguientes:

- R^2 : mide la cantidad de variación entre las variables dependientes; el modelo con mejor rendimiento tendrá un R^2 más cercano a 1
- MAPE: mide la diferencia promedio entre los valores reales y los valores predichos. Valores más bajos de MAPE indican un mejor rendimiento del modelo.(Carranza et al., 2022)
- nRMSE (Root Mean Square Error): mide la desviación promedio de los valores estimados(Jafary et al., 2024)

Recomendaciones

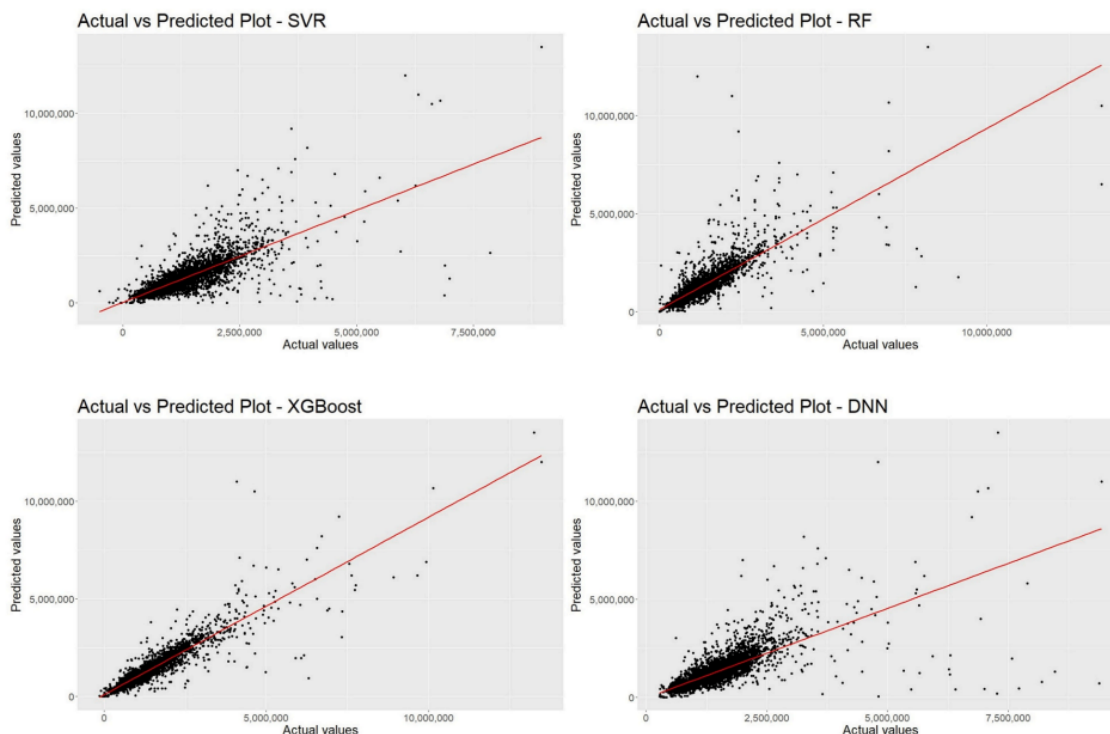
Results of the four methods considered for mass land valuation.

Method	SVR	RF	XGBoost	DNN
Training data				
R ²	0.674	0.996	0.999	0.647
MAPE	0.25	0.028	0.046	0.593
nRMSE	0.766	0.076	0.037	0.776
Test data				
R ²	0.649	0.701	0.862	0.611
MAPE	0.477	0.166	0.139	0.573
nRMSE	0.436	0.406	0.281	0.48

(Jafary et al., 2024)

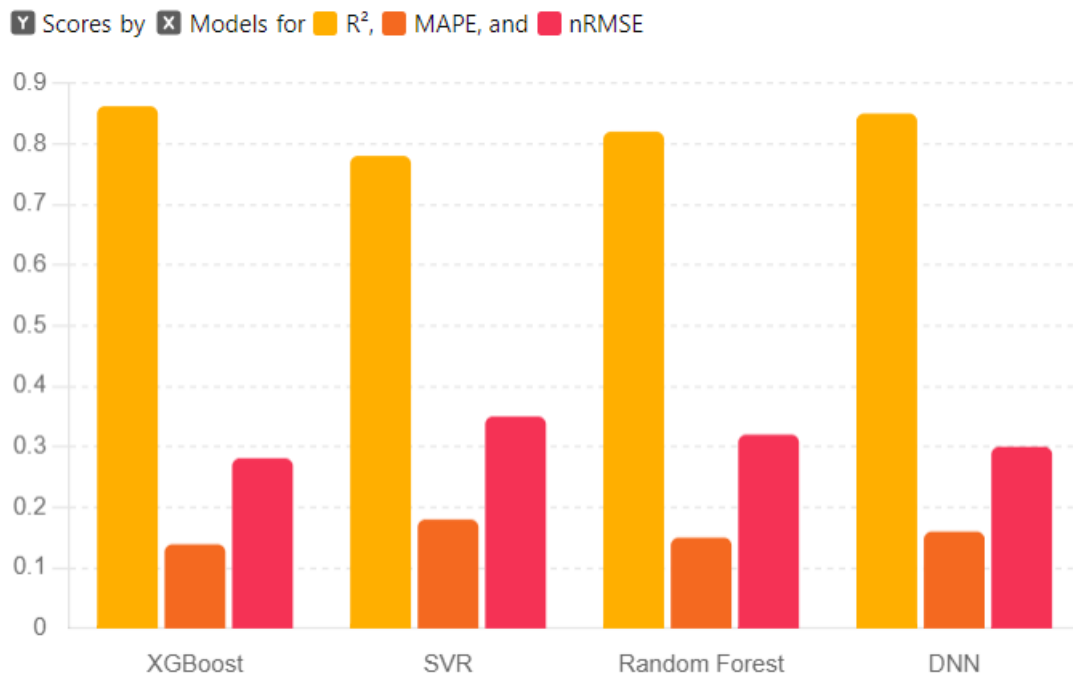
A partir de los datos presentados, podemos ver que tanto los modelos RF como XGBoost han superado considerablemente a los otros dos modelos. En el entrenamiento, el estudio encuentra que el modelo XGBoost obtiene el puntaje R^2 más alto. RF obtiene un R^2 comparable y un rendimiento superior según la métrica MAPE, aunque esto se compensa con un mejor nRMSE del modelo XGBoost. En los datos de prueba, el modelo XGBoost supera claramente a todos los demás modelos en todas las métricas utilizadas.

La superioridad de los modelos XGBoost y RF se muestra aún más en los gráficos de dispersión a continuación.



Podemos ver, al comparar los gráficos de dispersión, que tanto RF como XGBoost tienen pendientes que están mucho más cerca de 1 que los gráficos de dispersión de SVR y DNN. El rendimiento de los modelos muestra claramente que los modelos RF y

XGBoost han demostrado la capacidad de proporcionar valoraciones que están cerca de la realidad. Además, aunque esto no es un factor determinante del rendimiento, el modelo XGBoost demostró capacidades de computación más rápidas, lo que confirma aún más la superioridad del modelo XGBoost.



Al comparar los modelos SVR, Random Forest, XGBoost y DNN, hay una clara superioridad del Modelo XGBoost, tanto en precisión como en velocidad. Los otros modelos, aunque adecuados para valoraciones automatizadas, han mostrado un rendimiento inferior, con mayores tasas de error. La comparación de estos AVMs muestra que ciertos programas de aprendizaje automático son más adecuados para la valoración. Es importante notar que el modelo XGBoost es más complejo que sus competidores. Aunque esto no garantiza un mejor rendimiento, se puede argumentar que es un factor determinante en la brecha de rendimiento. En general, es el programa mejor adaptado de los cuatro para la tarea de valoraciones automatizadas de terrenos.

Caso práctico de AVM – Modelo 1 – Regresión Lineal

Introducción

Este caso tiene como objetivo modelar las ventas de casas en el condado de King, estado de Washington, EE. UU., utilizando la Regresión Lineal Múltiple (MLR).

El conjunto de datos, obtenido de Kaggle, comprende datos históricos de casas vendidas entre mayo de 2014 y mayo de 2015. Cubre una amplia región, incluyendo Seattle, y contiene 21 variables con 21,613 observaciones. Para este estudio, utilizamos los primeros 500 puntos de datos para el entrenamiento del modelo y los siguientes 100 para las pruebas. Nuestro objetivo es lograr una precisión de predicción del 75-80% e identificar los factores clave que contribuyen a valores de propiedad más altos, particularmente aquellos valorados en \$650,000 y más. (*Kc_house_data*, n.d.)

Data

El conjunto de datos incluye varias características relacionadas con las casas:

- **Price:** Precio de venta de la casa (variable dependiente).
- **Bedrooms:** Número de dormitorios.
- **Bathrooms:** Número de baños.
- **Sqft Living:** Metros cuadrados del espacio habitable.
- **Sqft Lot:** Metros cuadrados del terreno
- **Floors:** Número de pisos en la casa
- **Condition:** Condición de la casa, valorada en una escala
- **Grade:** Calificación general dada a la unidad de vivienda, basada en el sistema de calificación del condado de King
- **Yr Built:** Año en que se construyó la casa
- .

Otras características incluidas no se utilizaron en el modelo de regresión debido a problemas de redundancia o multicolinealidad.

Los datos fueron limpiados y preparados para el análisis de regresión en Microsoft Excel.

Metodología

Realizamos un análisis de Regresión Lineal Múltiple utilizando las primeras 500 observaciones para entrenar el modelo. La variable dependiente (Y) fue el precio de la casa, mientras que las variables independientes (Xs) incluyeron el número de dormitorios, baños, metros cuadrados del espacio habitable, tamaño del terreno, número de pisos, condición, calificación, tamaño del sótano y el año de construcción.

Fórmula del Modelo de Regresión:

$$\text{Price} = \beta_0 + \beta_1 \times \text{bedrooms} + \beta_2 \times \text{bathrooms} + \beta_3 \times \text{sqft_living} + \beta_4 \times \text{sqft_lot} + \beta_5 \times \text{floors} + \beta_6 \times \text{condition} + \beta_7 \times \text{grade} + \beta_8 \times \text{sqft_basement} + \beta_9 \times \text{yr_built}$$

Donde:

- β_0 es el “intercept”.
- $\beta_1, \beta_2, \dots, \beta_9$ son los coeficientes para cada variable predictora, estos coeficientes son proporcionados por el modelo de regresión

Salida y Análisis de la Regresión

La salida de la regresión incluye varias estadísticas y métricas clave que proporcionan información sobre el rendimiento del modelo:

<i>Regression Statistics</i>	
Multiple R	0.756475606
R Square	0.572255342
Adjusted R Square	0.565285979
Standard Error	230810.5403
Observations	500

Regression:

Multiple R: 0.75649. Este coeficiente de correlación indica una fuerte relación positiva entre los precios observados y los predichos.

R Square (R^2): 0.572277. Este valor sugiere que aproximadamente el 57.23% de la variabilidad en los precios de las viviendas puede ser explicada por el modelo. Es una medida del poder explicativo del modelo.

Adjusted R Square: 0.5652859. El R^2 ajustado tiene en cuenta el número de predictores en el modelo, proporcionando una medida más precisa al comparar modelos con un número diferente de predictores.

Standard Error: 230810.5403. Esto representa la distancia promedio que los valores observados caen desde la línea de regresión. Un error estándar más bajo indica un mejor ajuste.

Observations: 500. Esto indica el tamaño de la muestra utilizado en el análisis de regresión.

ANOVA (Analysis de Variancia):

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	8	3.5E+13	4.37E+12	82.11013514	1.36493E-85
Residual	491	2.62E+13	5.33E+10		
Total	499	6.12E+13			

El estadístico F: 82.11013514 con un valor de significancia de 1.36×10^{-85} indica que el modelo en general es estadísticamente significativo. El alto valor del estadístico F sugiere que el modelo explica una parte significativa de la varianza en los precios de las viviendas. El ANOVA, aunque no nos proporciona información orientada a la aplicación, confirma la validez estadística del modelo.

Tabla de coeficientes:

La tabla de coeficientes proporciona información crucial sobre cómo cada variable predictora influye en los precios de las viviendas. Cada coeficiente representa el cambio esperado en la variable dependiente (precio de la vivienda) por un cambio de una unidad en la variable predictora, manteniendo todas las demás variables constantes. La significancia de cada predictor se determina por su valor p, que indica si la relación observada podría haber ocurrido por azar.

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	6812003.649	911433.2491	7.473946836	3.59809E-13
bedrooms	-39261.28847	15248.08933	-2.574833319	0.010320561
bathrooms	52978.14244	25148.86791	2.106581602	0.035660652
sqft_living	163.9191081	23.27444537	7.042879241	6.39086E-12
sqft_lot	0.403500362	0.370807943	1.088165368	0.277055949
floors	-18057.83078	24510.03669	-0.736752499	0.461624749
condition	-13827.36013	15677.81742	-0.881969713	0.378224743
grade	124824.7921	15594.91455	8.004198531	8.75379E-15
yr_built	-3808.337879	470.8960122	-8.087428605	4.80183E-15

El intercept (β_0) es 6812003.649 con un valor p < 0.001 , lo que indica que es estadísticamente significativo. El intercepto representa el precio base de la vivienda

cuando todos los demás predictores son cero. Con un valor p altamente significativo, este valor es estadísticamente significativo, lo que indica que proporciona un punto de partida significativo para el modelo.

Bedrooms (β_1): -39261.28847 (p-value: 0.010320561). Esto sugiere que, manteniendo todas las demás variables constantes, un dormitorio adicional disminuye el precio en aproximadamente \$39261. El valor p es menor que 0.05, lo que indica significancia estadística. Esta relación negativa puede parecer contraintuitiva, sugiriendo una posible multicolinealidad con otras variables como el total de pies cuadrados o el número de baños.

Bathrooms (β_2): 52978.14244, p-value: 0.035660652. Cada baño adicional aumenta el precio de la vivienda en aproximadamente \$52,978.14. El valor p está por debajo de 0.05, lo que hace que este predictor sea estadísticamente significativo. Esta relación positiva se alinea con la expectativa de que más baños típicamente añaden valor a una propiedad.

Sqft Living (β_3): 163.9191081, p-value: 6.39e10-12. Cada pie cuadrado adicional de espacio habitable aumenta el precio de la vivienda en aproximadamente \$163.92. El valor p extremadamente bajo indica un nivel muy alto de significancia estadística, lo cual es lógico ya que confirma que el espacio habitable es un determinante crucial del precio de la vivienda.

Sqft Lot (β_4): 0.403, p-value: 0.277. El coeficiente para el tamaño del lote es positivo pero muy pequeño, lo que sugiere un impacto mínimo en los precios de las viviendas. El valor p alto (mayor que 0.05) indica que este predictor no es estadísticamente significativo, lo que implica que las variaciones en el tamaño del lote no influyen significativamente en los precios de las viviendas dentro del contexto de este modelo. Esto sugiere que el tamaño del lote no tiene un impacto sustancial en los precios de las viviendas en este modelo.

Floors (β_5): -18057.83078 (p-value: 0.4616). No es estadísticamente significativo debido al valor p que es demasiado alto para considerar este punto de datos relevante. El número de pisos no afecta significativamente el precio de la vivienda, podríamos suponer que esto se debe a que la superficie habitable compensa este punto de datos.

Condition (β_6): -13827.36013 (p-value: 0.378). La condición de la casa no tiene un impacto estadísticamente significativo en su precio, al menos para el conjunto de datos probado. Esto puede explicarse si la mayoría de las propiedades en el conjunto de datos estuvieran en una condición similar. Es importante tener en cuenta que el puntaje promedio de la propiedad no está determinado por este punto de datos, sino por la "calidad" que se le atribuye.

Grade (β_7): 124824.7921 (p-value: $8.75379 \times 10^{-15} < 0.001$). Cada aumento de una unidad en la calificación de grado aumenta el precio de la vivienda en aproximadamente \$124,824.79. El valor p muy bajo indica una alta significancia estadística, confirmando que las calificaciones más altas, indicativas de una mejor

calidad de construcción y diseño, aumentan significativamente los precios de las viviendas.

Year Built (β_9): -3808.337879 (p-value: $4.80183 \times 10^{-15} < 0.001$). El coeficiente indica que las casas más antiguas tienden a tener precios más bajos, con cada año adicional reduciendo el precio en aproximadamente \$3808.34. El valor p muy bajo demuestra que esta relación es estadísticamente significativa, probablemente reflejando depreciación o el atractivo anticuado de las propiedades más antiguas.

Evaluación del Modelo

El valor de R^2 del modelo de 0.572255 indica que explica más del 57% de la variabilidad en los precios de las viviendas. Este es un nivel razonable de precisión, pero también sugiere que casi el 43% de la variabilidad se explica por factores no incluidos en el modelo.

Predictores significativos como el número de baños, el espacio habitable en pies cuadrados y la calidad tienen un claro impacto en los precios de las propiedades. Por otro lado, variables como el tamaño del lote, el número de pisos, la condición y el tamaño del sótano no fueron estadísticamente significativas. Esto puede deberse a las características específicas del mercado de viviendas en el condado de King o a la multicolinealidad entre variables.

ID	Actual Price	Model Price	% error	ID	Actual Price	Model Price	% error	ID	Actual Price	Model Price	% error	ID	Actual Price	Model Price	% error
9485740340	2125000	420518	-80%	5379805885	529000	271327	-48.71%	2.77E+09	442000	411018	-7%	7.846E+09	515000	798800	55%
561000075	671500	148800	-78%	7852030960	650000	355489	-45.31%	7.943E+09	900000	865107	-4%	4.377E+09	331500	517106	56%
1521069070	358000	90009.2	-75%	1922059046	750000	410226	-45.30%	1.715E+09	650000	649172	0%	8.906E+09	240000	376756	57%
8945200860	619000	169875	-73%	7511800070	730000	404747	-44.56%	1.311E+09	530000	544185	3%	2.725E+09	638700	1037176	62%
7453000070	727500	201902	-72%	9238901420	615000	352742	-42.64%	7.203E+09	260000	279602	8%	2.525E+09	265000	431458	63%
7202331420	1600000	456451	-71%	7800800160	910000	523633	-42.46%	8.121E+09	480000	529957	10%	224069094	375000	625336	67%
925059288	2250000	648537	-71%	3278600240	595000	374273	-37.10%	5.307E+09	435000	481472	11%	4.025E+09	261000	458032	75%
8669160310	930000	269137	-71%	4348800030	650000	412907	-36.48%	4.403E+09	308000	344017	12%	3.886E+09	585000	1049703	79%
5113400431	1485000	431048	-71%	8018600765	372500	239507	-35.70%	1.123E+09	240500	287626	20%	1.546E+09	200000	402949	101%
1843200240	970000	299928	-69%	9542830350	360000	237282	-34.09%	3.624E+09	404950	504715	25%	2.473E+09	215000	448776	109%
1180003090	515000	181958	-65%	2270000070	318888	212948	-33.22%	2.781E+09	296000	373235	26%	726049331	235000	494748	111%
5104510270	720000	257715	-64%	2459500310	452000	302813	-33.01%	222069082	266000	341662	28%	98020350	442000	933263	111%
4309710240	1950000	722250	-63%	8648100200	460000	314699	-31.59%	7.9E+09	315000	413776	31%	1.526E+09	530000	1278701	141%
4099500935	1450000	537780	-63%	662310400	1220000	846266	-30.63%	2.008E+09	204000	271566	33%	1.089E+09	180000	459004	155%
6381500700	662000	248398	-62%	1921069084	880000	622401	-29.27%	6.909E+09	505000	687494	36%	7.853E+09	264000	685231	160%
236400320	750000	285016	-62%	5466410030	346900	245968	-29.10%	2.944E+09	782000	1072831	37%	5.695E+09	249000	653134	162%
5000500055	705000	282390	-60%	5419800510	330000	239551	-27.41%	2.999E+09	439000	603208	37%	339500160	270500	719778	166%
8718500075	725000	327593	-55%	1442740140	685000	512279	-25.21%	3.62E+09	290000	411659	42%	1.374E+09	451000	1282790	184%
5126310400	1095000	502019	-54%	6431500140	863000	675029	-21.78%	4.387E+09	890000	1270023	43%	1.245E+09	280000	802390	187%
192460060	640000	295485	-54%	3340401555	275000	216574	-21.25%	3.876E+09	268500	390604	45%	4.139E+09	259950	755833	191%
5467900070	1197350	559399	-53%	1432400335	190000	150341	-20.87%	7.519E+09	342000	500192	46%	7.504E+09	370000	1101767	198%
3294700310	396000	188658	-52%	4045500710	338900	268256	-20.85%	809002765	405000	609524	50%	1.189E+09	263500	840331	219%
6204420070	595000	285892	-52%	2436701180	610000	555949	-8.86%	9.276E+09	300000	456907	52%	1.923E+09	288000	961724	234%
2827100070	615000	312005	-49%	2473360060	300000	278862	-7.05%	7.852E+09	475000	728585	53%	7.323E+09	290000	1000435	245%
5379805885	529000	271327	-49%	2769600560	442000	411018	-7.01%	7.846E+09	515000	798800	55%	3.225E+09	365000	1601096	339%
7852030960	650000	355489	-45%	7942601895	900000	865107	-3.88%	4.377E+09	331500	517106	56%	5.487E+09	238000	1301421	447%
1922059046	750000	410226	-45%	1714900060	650000	649172	-0.13%	8.906E+09	240000	376756	57%	622049114	280000	1603229	473%
7511800070	730000	404747	-45%	1310900260	530000	544185	2.68%	2.725E+09	638700	1037176	62%	4.404E+09	160000	1238356	674%

Al probar el modelo para 100 valores siguiendo la fórmula del modelo, podemos ver claramente que el modelo puede considerarse altamente inexacto. Esto ilustra los datos de R^2 , ya que podemos ver que el modelo está considerablemente limitado en su capacidad para proporcionar una estimación precisa del valor de la propiedad. Esto se debe a las limitaciones del modelo, como se muestra por el R^2 , el modelo solo puede explicar el 57% de los precios y, por lo tanto, hay grandes inexactitudes en las estimaciones.

Limitaciones y Mejoras

El modelo de regresión actual presenta varias limitaciones que deben abordarse para mejorar su precisión predictiva y confiabilidad. Estas limitaciones incluyen la multicolinealidad, variables no significativas, complejidad del modelo y la posible aplicación de técnicas avanzadas.

La multicolinealidad ocurre cuando dos o más variables predictoras en un modelo de regresión están altamente correlacionadas, lo que significa que proporcionan información redundante sobre la variable de respuesta. Esto puede aumentar los errores estándar de los coeficientes, lo que dificulta determinar el efecto individual de cada predictor. En nuestro modelo, el coeficiente negativo para el número de dormitorios sugiere una posible multicolinealidad con otras variables como el total de pies cuadrados o el número de baños.

Por ejemplo, una casa más grande a menudo tiene más dormitorios y más espacio habitable en general. Si tanto el número de dormitorios como el total de pies cuadrados se incluyen en el modelo, puede ser difícil aislar el efecto de uno del otro. Esto puede resultar en estimaciones de coeficientes engañosas.

Las variables no significativas son una fuente de error. En nuestro modelo, variables como el tamaño del lote, el número de pisos y la condición no fueron estadísticamente significativas, lo que significa que sus valores p fueron mayores que 0.05. Estas variables no contribuyen de manera significativa al modelo y podrían ser transformadas o excluidas para mejorar el rendimiento del modelo.

La complejidad del modelo también es un factor que se puede utilizar para aumentar la complejidad. Agregar más puntos de datos y explorar otros posibles predictores podría mejorar la capacidad predictiva del modelo. El modelo actual utiliza una muestra relativamente pequeña de 500 observaciones de un posible total de 21,613. Aumentar el tamaño de la muestra puede mejorar la robustez y generalizabilidad del modelo. Más puntos de datos proporcionan una imagen más clara de los patrones subyacentes y reducen el impacto de valores atípicos o anomalías. Además, explorar otros posibles predictores podría mejorar la capacidad predictiva del modelo. Variables como la proximidad a servicios (distancia a escuelas, centros comerciales y transporte público), características del vecindario (tasas de criminalidad, niveles de ingresos promedio y calidad de las escuelas locales ...) o factores temporales (tendencias del mercado inmobiliario a lo largo del tiempo, estacionalidad e indicadores económicos). Al

incorporar estos predictores adicionales, podemos desarrollar un modelo más completo que capture mejor las complejidades del mercado de la vivienda.

En conclusión, el modelo de regresión lineal múltiple proporciona una predicción razonable de los precios de las viviendas en el condado de King, explicando más del 57% de la varianza. Predictores significativos como los baños, los pies cuadrados y la calificación juegan un papel crucial en la determinación de los precios de las viviendas. Para mejorar el rendimiento del modelo, sería clave centrarse en abordar la multicolinealidad, mejorar la significación de las variables, aumentar la complejidad del modelo y explorar predictores adicionales para mejorar la precisión y confiabilidad del modelo.

Caso Práctico de AVM - Modelo 2 - Regresión Lineal

Introducción

Este segundo modelo es una mejora respecto al primero. La diferencia aquí radica en que el tamaño de la muestra es mucho más grande, vamos a utilizar un total de 21611 observaciones, abarcando numerosas variables como dormitorios, baños, pies cuadrados, pisos, presencia de frente al agua, vista, calidad, año de construcción y año de renovación. El objetivo es producir un AVM más adecuado que el anterior al proporcionar más puntos de datos y observaciones para permitir que el modelo prediga mejor los precios.

Idoneidad del modelo

<i>Regression Statistics</i>	
Multiple R	0.808543209
R Square	0.653742122
Adjusted R Square	0.653517654
Standard Error	216110.4992
Observations	21611

Al interpretar el valor de R-cuadrado (R^2) de 0.653 de mi análisis de regresión, entiendo que R^2 representa la proporción de la varianza en la variable dependiente (precios de las viviendas) que es explicada por las variables independientes (predictores) incluidas en mi modelo. Un valor de R^2 de 0.653 significa que el 65.3% de la variabilidad en los precios de las viviendas es explicada por los predictores en mi modelo. Esto indica que mi modelo tiene un poder explicativo sólido, pero también sugiere que todavía hay un 34.7% de la varianza en los precios de las viviendas que no es explicada por los predictores que he elegido. Esta varianza no explicada podría deberse a factores que no están incluidos en mi modelo o a la aleatoriedad inherente en los datos.

Para refinar aún más esta medida, observo el valor de R-cuadrado ajustado, que es de 0.653. Una mejor medida de si mi modelo se ajusta a los datos es proporcionada por el R-cuadrado ajustado, que modifica el valor de R^2 para tener en cuenta el número de predictores en mi modelo. Esta modificación es crucial porque aumentar el número de predictores en el modelo tiene el potencial de aumentar artificialmente el valor de R-cuadrado sin mejorar la capacidad predictiva real del modelo. Con un R-cuadrado ajustado de 0.653, mi modelo explica el 65.3% de la varianza en los precios de las

viviendas después de ajustar por el número de predictores. Esto implica que mi modelo es fuerte y a partir de ahora, probaremos su significancia.

Significancia del Modelo

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	14	1.90428E+15	1.3602E+14	2912.407143	0
Residual	21596	1.00861E+15	46703747853		
Total	21610	2.9129E+15			

Para determinar si existe una asociación significativa entre el conjunto de variables independientes (predictores mencionados anteriormente) y la variable dependiente (precios de las viviendas), obtengo un Estadístico F de 2912.407 en mi análisis. El modelo de regresión con predictores se compara con un modelo sin predictores (solo el intercepto) utilizando el Estadístico F. Con un Estadístico F general de 2912.407, mis predictores juntos explican una parte significativa de la varianza en los precios de las viviendas, lo que sugiere que mi modelo se ajusta considerablemente mejor a los datos que un modelo sin ningún predictor.

El valor de Significancia F, que es 0, valida aún más la importancia de mi modelo. La hipótesis nula, que establece que ningún cambio en la variable dependiente puede ser explicado por los predictores, se prueba mediante este valor p. Todos los coeficientes de regresión deben ser iguales a cero. Con un valor p de cero o casi cero, es muy improbable que la correlación observada entre los predictores y los precios de las viviendas sea el resultado del azar. Por lo tanto, puedo afirmar con confianza que el modelo de regresión es estadísticamente significativo.

Significancia de las Variables Independientes

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	6195562.082	138441.0394	44.75235168	0	5924207.423	6466916.742	5924207.423	6466916.742
bedrooms	-39329.22854	2026.924755	-19.4033984	3.68931E-83	-43302.15073	-35356.30636	-43302.15073	-35356.30636
bathrooms	45734.417	3492.602086	13.09465432	4.97589E-39	38888.65903	52580.17498	38888.65903	52580.17498
sqft_living	167.1254538	4.669115719	35.79381277	1.2355E-272	157.9736423	176.2772654	157.9736423	176.2772654
floors	26873.84721	3784.017055	7.101936068	1.26839E-12	19456.89437	34290.80004	19456.89437	34290.80004
waterfront	579078.966	18636.96289	31.07153077	2.1225E-207	542549.1426	615608.7894	542549.1426	615608.7894
view	43218.97102	2273.307372	19.01149468	6.1636E-80	38763.12071	47674.82133	38763.12071	47674.82133
grade	119714.1721	2249.240013	53.22427636	0	115305.4956	124122.8486	115305.4956	124122.8486
sqft_above	-6.298103683	4.545104923	-1.385689393	0.165856078	-15.20684494	2.61063757	-15.20684494	2.61063757
sqft_living15	24.86049176	3.60050878	6.904716326	5.16995E-12	17.80322869	31.91775482	17.80322869	31.91775482
sqft_lot15	-0.550602481	0.078363065	-7.026301012	2.18387E-12	-0.704199874	-0.397005088	-0.704199874	-0.397005088
sqft_lot	-0.001646101	0.051285073	-0.032097087	0.974394923	-0.102168632	0.098876429	-0.102168632	0.098876429
condition	19493.28799	2496.928808	7.806905802	6.1265E-15	14599.12316	24387.45283	14599.12316	24387.45283
yr_built	-3570.202071	70.99416398	-50.28866981	0	-3709.355874	-3431.048267	-3709.355874	-3431.048267
yr_renovated	10.14758331	3.913989545	2.592644459	0.009530535	2.475874796	17.81929182	2.475874796	17.81929182

Examinando los predictores individuales, considerando el coeficiente de cada predictor, el valor p y si es significativo al nivel de 0.05.

Bedrooms: El coeficiente para los dormitorios es -39,329.23, lo que significa que un aumento de un dormitorio está vinculado a una caída de aproximadamente \$39,329.23 en el precio de la vivienda, manteniendo todas las demás cosas iguales. Un predictor estadísticamente significativo de los precios de las viviendas, el número de dormitorios tiene un valor p de 3.69E-81, que es menor que 0.05. Esto indica que el precio de una casa está influenciado significativamente por el número de dormitorios.

Bathrooms: El coeficiente para los baños es 45,734.42, lo que significa que un baño adicional está relacionado con un aumento de \$45,734.42 en el precio de la vivienda, manteniendo todas las demás cosas constantes. Los baños tienen un valor p de 4.98E-39, que es menor que 0.05 y esto indica que son un predictor significativo de los precios de las viviendas.

Sqft_living: El coeficiente para sqft_living es 167.13, lo que indica un aumento de \$167.13 en el precio de la vivienda por cada pie cuadrado de área habitable adicional. Un predictor muy importante del precio de la propiedad es el pies cuadrados de espacio habitable, como lo demuestra el valor p muy bajo de 1.97E-272.

Floors: El coeficiente relacionado con los pisos es de 26,873.85, lo que indica que un piso adicional corresponde a un aumento de \$26,873.85 en el precio de una propiedad. El número de pisos es un predictor significativo del precio de la propiedad, como lo indica el valor p de 1.26E-12.

Waterfront: El coeficiente para la propiedad frente al agua es de 57,097.90, lo que indica que las propiedades ubicadas frente al agua tienen un precio \$57,097.90 más alto, manteniendo todas las demás cosas iguales. El valor p de 2.12E-207 muestra una alta significancia.

View: El coeficiente para la vista es de 43,218.97, lo que significa que una propiedad con vista tiene un valor \$43,218.97 más alto, manteniendo constantes otras variables. El valor p de 6.16E-38 significa su importancia.

Grade: El coeficiente para la calidad es de 119,714.17, lo que indica que una calidad más alta se corresponde con un aumento de \$119,714.17 en el precio de una casa. Como lo demuestra el valor p extremadamente bajo de 2.26E-296, la calidad es un predictor altamente significativo del precio de la vivienda.

Yr_built: El coeficiente para el año de construcción es de -3,570.20, lo que indica que el precio de una casa disminuye alrededor de \$3,570.20 por cada año que ha pasado desde que se construyó. Dado el valor p extremadamente bajo de 6.13E-15, se puede concluir que el año de construcción es un predictor altamente significativo del precio.

Matriz de Correlación

La matriz de correlación ofrece información importante sobre los vínculos entre los predictores y los precios de las viviendas. Los coeficientes de correlación en la diagonal demuestran el grado de relación entre cada predictor individual y los precios de las viviendas.

	price	bedrooms	bathrooms	sqft_living	floors	waterfront	view	grade	sqft_above	sqft_living15	sqft_lot15	sqft_lot	condition	yr_built	yr_renovated
price	1														
bedrooms	0.308366	1													
bathrooms	0.52515	0.515974	1												
sqft_living	0.702055	0.576763	0.75468405	1											
floors	0.256814	0.175418	0.50071194	0.354048	1										
waterfront	0.266371	-0.006581	0.06374329	0.103829	0.0237	1									
view	0.397299	0.079537	0.18773462	0.284647	0.0294	0.401857	1								
grade	0.667434	0.356998	0.66498055	0.762727	0.4582	0.082775	0.2513	1							
sqft_above	0.605567	0.477616	0.68536337	0.876644	0.5239	0.072074	0.1676	0.7559	1						
sqft_living15	0.585377	0.39167	0.56862617	0.75644	0.2799	0.086463	0.2804	0.7132	0.731871	1					
sqft_lot15	0.082448	0.029252	0.08716281	0.183301	-0.0113	0.030702	0.0726	0.1192	0.194048	0.18318821	1				
sqft_lot	0.089661	0.03171	0.08773009	0.172841	-0.0052	0.021602	0.0747	0.1136	0.183511	0.14460512	0.718556	1			
condition	0.036379	0.028433	-0.1249169	-0.058689	-0.2638	0.016655	0.046	-0.1446	-0.15821	-0.0927952	-0.0034	-0.00895	1		
yr_built	0.054006	0.154248	0.50596774	0.318066	0.4894	-0.02617	-0.053	0.447	0.423915	0.32621424	0.070936	0.053061	-0.3614	1	
yr_renovated	0.126436	0.018844	0.05073262	0.055377	0.0063	0.092883	0.1039	0.0144	0.023283	-0.0026748	0.007849	0.00764	-0.0606	-0.224913	1

Se observan correlaciones fuertes entre los precios de las viviendas y factores como el espacio habitable en pies cuadrados (sqft_living), la calidad y el espacio sobre el suelo (sqft_above), todos los cuales muestran correlaciones positivas bastante altas. Por ejemplo, el coeficiente de correlación entre el precio y sqft_living es de 0.702, lo que indica un efecto sustancial en los precios de las viviendas con cada aumento en el espacio habitable.

Otras correlaciones moderadamente positivas, como las entre vista y frente al agua, indican efectos interdependientes en los valores de las viviendas, cada uno de los cuales contribuye a la dinámica general del mercado inmobiliario.

Estas estadísticas, respaldadas por datos significativos, destacan la complejidad de las relaciones entre los predictores y los precios de las viviendas. Enfatizan lo importante que es considerar múltiples elementos simultáneamente para comprender completamente las complejidades del mercado inmobiliario y tomar decisiones informadas dentro de él.

Evaluación del Modelo

price	bedro		bathro		sqft_liv		waterf		sqft_abov		sqft_living		yr_renova		PRICE	ACCORDING	Difference	Difference	Average error%
	oms	oms	ing	ing	floors	ront	view	grade	e	15	sqft_lot15	sqft_lot	condition	yr_built	ted	TO MODEL	%	% Absolute	
525000	2	2	1340	2	0	0	7	1340	1340	3600	3600	3	1903	0	575034.1625	10%	10%	31%	
242000	3	2	1310	1	0	0	7	1310	1440	9893	9645	3	1979	0	256552.5938	6%	6%		
715000	4	3	2650	1	0	0	8	1680	2650	30500	30500	4	1960	0	698883.8875	-2%	2%	Standard Dev	
475000	3	3	1890	2	0	0	8	1890	1890	4142	3938	3	1999	0	485113.9822	2%	2%	45%	
370000	4	3	2580	2	0	0	8	2580	2010	5600	5511	3	1999	0	570367.1208	54%	54%		
788600	4	3	3500	2	0	0	9	3500	3920	7666	7200	3	2005	0	851530.8526	8%	8%	Variance	
160000	2	1	1040	1	0	0	6	1040	910	5080	13100	5	1912	0	366101.8696	129%	129%	20%	
800000	5	2	1930	2	0	0	7	1930	1710	87120	501376	3	1930	0	442802.3584	-45%	45%		
520000	4	3	2190	2	0	0	8	2190	2130	5142	5085	4	1992	0	555365.2929	7%	7%		
725000	3	2	1530	2	0	0	8	1530	2140	5000	4000	3	1985	0	437207.1521	-40%	40%		
335000	4	1	1480	2	0	0	7	1480	2380	6000	6500	4	1914	0	500774.083	49%	49%		
919000	4	3	3620	1	0	4	10	2220	3530	17026	17133	3	1993	0	1161906.928	26%	26%		
425000	3	3	3370	2	0	0	9	2650	2150	14048	13929	3	1986	0	894792.9076	111%	111%		
420000	5	3	2100	1	0	0	7	1140	1830	8700	14395	3	1983	0	341358.5689	-19%	19%		
905000	4	3	3370	2	0	0	10	3370	3370	38896	47959	4	1981	0	1036013.532	14%	14%		
435000	2	1	2240	1	0	0	7	1120	1390	6000	7200	4	1940	0	577839.9857	33%	33%		
438000	3	3	1650	2	0	0	7	1650	1560	3070	3031	3	2003	0	293474.6102	-33%	33%		
725000	4	3	2630	2	0	0	8	2630	2670	7506	7505	3	1994	0	600181.0883	-17%	17%		
840000	4	2	2330	2	0	0	8	1300	2040	4000	4000	5	1924	0	787860.3061	-6%	6%		
425500	3	3	1970	2	0	0	8	1970	1850	2778	2752	3	1978	0	561279.3639	32%	32%		
853505	3	3	2820	1	0	4	9	1770	3740	14890	14890	3	1985	0	985617.8666	15%	15%		
398000	2	1	590	2	0	0	5	590	2020	15180	10945	3	1983	0	-82981.0014	-121%	121%		
418500	3	3	2060	2	0	0	7	2060	2060	4399	4399	3	2007	0	356829.2588	-15%	15%		
826000	3	2	1890	2	0	0	9	1890	2560	5000	5000	3	1929	0	800319.7188	-3%	3%		

Para evaluar la precisión del nuevo Modelo de Valoración Automatizada (AVM), se realizó una prueba utilizando 100 conjuntos de datos aleatorios. Los resultados, como se muestra en la captura de pantalla, proporcionan una visión general de los precios reales, los precios predichos por el modelo y la diferencia porcentual entre estos valores. La precisión del modelo se evalúa en función de varios indicadores clave, incluido el porcentaje de error promedio, la desviación estándar y la varianza.

Métricas Claves

Average Error Percentage: 31%

Standard Deviation: 45%

Variance: 20%

Análisis

Average Error Percentage

El porcentaje de error promedio es un indicador crucial de la precisión del modelo. En este caso, el modelo muestra un porcentaje de error promedio del 31%, lo que indica que, en promedio, las predicciones del modelo se desvían de los precios reales en un 31%. Este nivel de error sugiere que el modelo tiene un margen considerable para mejorar en términos de precisión.

Standard Deviation

La desviación estándar del 45% refleja la dispersión de los errores de predicción. Una alta desviación estándar indica que hay una variación significativa en los errores de predicción, lo que significa que mientras algunas predicciones están muy cerca de los precios reales, otras están sustancialmente alejadas. Esta alta desviación estándar sugiere que el rendimiento del modelo es inconsistente en diferentes puntos de datos.

Variance

La varianza del 20% proporciona una medida de la dispersión de los errores de predicción alrededor de la media. Una varianza más alta indica una mayor variabilidad en la precisión del modelo. La varianza del 20% en este contexto señala un nivel moderado de variabilidad en las predicciones del modelo.

Observaciones

El modelo tiene un desempeño excepcionalmente bueno para ciertos puntos de datos, prediciendo precios con un error mínimo. Por ejemplo, la predicción del modelo para una casa con un precio de \$715,000 está fuera por solo un 2%.

Sin embargo, hay casos en los que las predicciones del modelo se desvían significativamente de los precios reales. Por ejemplo, una casa con un precio real de \$335,000 se predice en \$500,774.08, mostrando una diferencia del 49%.

El punto de datos con la desviación positiva más grande tiene un precio real de \$160,000, pero el modelo predice \$366,101.87, una diferencia del 129%. Este error significativo indica posibles problemas con el manejo de ciertas características de la propiedad o valores atípicos en los datos por parte del modelo.

Por otro lado, la mayor desviación negativa se observa para una propiedad con un precio real de \$800,000, pero el modelo predice \$442,802.36, lo que resulta en un error del -45%.

Conclusión

El análisis del nuevo modelo utilizando 100 conjuntos de datos aleatorios revela que el modelo muestra un porcentaje de error promedio del 31%, lo que indica una inexactitud sustancial en las predicciones. La alta desviación estándar y varianza indican un rendimiento inconsistente. Las predicciones del modelo son precisas para algunos puntos de datos pero significativamente incorrectas para otros, resaltando áreas donde el modelo podría ser mejorado. Esto sugiere que si bien el AVM es efectivo para predicciones generales, se necesita un mayor refinamiento para mejorar su precisión y confiabilidad, especialmente en casos que involucran propiedades con características únicas o extremas. La tasa de error promedio significativa destaca la necesidad de ajustes y optimización del modelo para mejorar su precisión general.

El AVM basado en Regresión Lineal Múltiple demuestra un fuerte poder explicativo con un valor de R-cuadrado de 0.653 y un R-cuadrado ajustado de 0.653, lo que indica que el modelo explica una parte significativa de la variación en los precios de las viviendas. El Estadístico F y los valores p de los predictores confirman la significancia general del modelo y la importancia de variables individuales como dormitorios, baños, espacio habitable, pisos, presencia frente al agua, vista, calidad y año de construcción. La matriz de correlación también respalda la validez del modelo al mostrar relaciones fuertes y moderadas entre los predictores y los precios de las viviendas. Este análisis exhaustivo subraya la capacidad del AVM en la valoración de propiedades, al tiempo que destaca áreas para posibles mejoras y refinamientos.

Comparación de Casos Prácticos de AVM - Modelo 1 y Modelo 2

El Modelo 1 utiliza 500 puntos de datos para entrenamiento y 100 para pruebas, con un R-cuadrado de 0.5723 y un R-cuadrado ajustado de 0.5653. Explica el 57.23% de la varianza en los precios de las viviendas. Los predictores significativos incluyen dormitorios, baños, pies cuadrados habitables, grado y año de construcción, mientras que los pies cuadrados de lote, pisos y condición no son significativos. El estadístico F del modelo es 82.11, con un nivel de significancia de 1.36×10^{-85} . Las pruebas muestran una alta inexactitud debido al 43% de la varianza no explicada, lo que indica multicolinealidad y variables no significativas.

El Modelo 2 mejora el Modelo 1 al utilizar 21,611 observaciones, logrando un R-cuadrado y un R-cuadrado ajustado de 0.653. Este modelo explica el 65.3% de la varianza en los precios de las viviendas con un estadístico F significativamente más alto de 2912.407. Los predictores significativos incluyen dormitorios, baños, pies cuadrados habitables, pisos, frente al agua, vista, grado y año de construcción. El porcentaje de error promedio es del 31%, con una desviación estándar del 45% y una varianza del 20%, lo que indica un amplio margen de mejora a pesar de un mejor rendimiento que el Modelo 1.

En general, el Modelo 2 demuestra un poder explicativo superior debido a una gama más amplia de predictores significativos y un conjunto de datos más grande. Sin embargo, ambos modelos requieren un mayor refinamiento para mejorar la precisión y confiabilidad.

Ventajas y Desventajas de los AVMs

Ventajas de los AVMs

Velocidad y Eficiencia

Una de las ventajas más significativas de los AVMs es su capacidad para proporcionar valoraciones rápidas de propiedades. Las valoraciones tradicionales de propiedades a menudo requieren varios días a semanas, involucrando múltiples pasos como visitas al sitio, recopilación de datos y preparación de informes. En contraste, los AVMs pueden procesar grandes volúmenes de datos y generar una valoración de propiedad en cuestión de minutos. Este proceso rápido es especialmente beneficioso en mercados inmobiliarios acelerados donde las decisiones oportunas son cruciales para compradores, vendedores y prestamistas. Por ejemplo, los prestamistas hipotecarios pueden agilizar las aprobaciones de préstamos y los agentes de bienes raíces pueden proporcionar a los clientes estimaciones de precios inmediatas, mejorando la eficiencia general de la transacción.

Por ejemplo, en el proceso de préstamos hipotecarios, un AVM puede utilizarse para determinar rápidamente el valor de una propiedad cuando un prestatario solicita un préstamo. Esto permite al prestamista tomar decisiones más rápidas sobre las aprobaciones de préstamos, lo cual es crucial en mercados competitivos donde los compradores necesitan actuar rápidamente.

Rentabilidad

Los AVM ofrecen una alternativa rentable a las valoraciones de propiedades tradicionales. La automatización del procesamiento de datos y la reducción de la necesidad de intervención humana disminuyen significativamente el costo de las valoraciones. Las valoraciones tradicionales implican tarifas para tasadores profesionales, costos administrativos y, a veces, gastos de viaje para visitas al sitio. Los AVM eliminan muchos de estos gastos, lo que los convierte en una opción atractiva para las valoraciones frecuentes necesarias para prestamistas, inversores inmobiliarios y gestores de cartera. Esta ventaja de costo es particularmente pronunciada en operaciones a gran escala, como la refinanciación de hipotecas o la gestión de carteras de propiedades, donde se requieren cientos o miles de valoraciones.

Por ejemplo, los fideicomisos de inversión inmobiliaria (REITs) a menudo gestionan grandes carteras de propiedades. Utilizando AVMs, pueden realizar valoraciones regulares de sus activos a una fracción del costo de las valoraciones tradicionales, lo que les ayuda a tomar decisiones de inversión informadas sin incurrir en gastos significativos.

Consistencia y Objetividad

Otro beneficio clave de los AVMs es su consistencia y objetividad. Las valoraciones tradicionales pueden verse afectadas por sesgos humanos e inconsistencias, ya que diferentes tasadores pueden tener opiniones variadas sobre la misma propiedad. Sin embargo, los AVMs aplican algoritmos estandarizados para evaluar los valores de las propiedades, asegurando que cada propiedad sea evaluada utilizando los mismos criterios. Esta uniformidad reduce el riesgo de errores de juicio subjetivo y proporciona un punto de referencia más confiable para los valores de las propiedades. Por ejemplo, en escenarios de préstamos, las valoraciones consistentes de las propiedades garantizan un trato justo a los prestatarios y reducen el riesgo de discrepancias en los montos de los préstamos.

Por ejemplo, durante el proceso de suscripción de líneas de crédito con garantía hipotecaria (HELOCs), los prestamistas pueden utilizar AVMs para garantizar que todas las propiedades sean valoradas de manera consistente, proporcionando una base equitativa para las decisiones de préstamos.

Perspectivas Basadas en Datos

Los AVMs aprovechan amplios conjuntos de datos para producir sus valoraciones. Estos conjuntos de datos típicamente incluyen datos de ventas históricas, atributos de propiedad (como tamaño, antigüedad y condición) e información del vecindario (como calidad de las escuelas, tasas de criminalidad y proximidad a servicios). Al analizar estos múltiples puntos de datos, los AVMs pueden identificar patrones y tendencias que influyen en los valores de las propiedades. Este enfoque basado en datos puede mejorar la precisión y confiabilidad de las valoraciones. Además, los AVMs pueden actualizarse continuamente con nuevos datos, lo que les permite reflejar las condiciones y tendencias del mercado actual con mayor precisión que los métodos tradicionales.

Por ejemplo, plataformas inmobiliarias en línea como Zillow y Redfin utilizan AVMs para proporcionar a los usuarios estimaciones en tiempo real de los valores de las viviendas. Estas estimaciones ayudan a los compradores y vendedores a comprender las condiciones del mercado y tomar decisiones informadas basadas en información actualizada.

Desventajas de los AVMs

Precisión Limitada en Propiedades Únicas

A pesar de sus fortalezas, los AVMs pueden enfrentar dificultades con propiedades únicas o no convencionales. Las propiedades con características arquitectónicas distintivas, importancia histórica o aquellas ubicadas en áreas con datos limitados de ventas comparables pueden presentar desafíos para los AVMs. Por ejemplo, una mansión histórica con características únicas puede no tener suficientes ventas comparables en la base de datos, lo que lleva a una valoración menos precisa. En tales casos, la evaluación detallada y matizada proporcionada por un tasador humano, que puede considerar las características únicas de la propiedad y las condiciones del mercado local, suele ser más confiable.

Por ejemplo, una casa construida a medida con acabados de alta gama y detalles arquitectónicos únicos puede no ser valorada con precisión por un AVM porque carece de propiedades comparables en la base de datos. Un tasador humano sería capaz de tener en cuenta estas características únicas y proporcionar una valoración más precisa.

Calidad y Disponibilidad de Datos

La fiabilidad de la valoración de un AVM depende en gran medida de la calidad y la completitud de los datos que utiliza. En regiones donde los datos inmobiliarios son escasos, desactualizados o inexactos, la valoración del AVM puede comprometerse significativamente. Por ejemplo, áreas rurales o mercados con bajos volúmenes de transacciones pueden no tener datos suficientes para que el AVM genere valoraciones precisas. Además, las discrepancias en las fuentes de datos pueden llevar a inconsistencias en las salidas del AVM. Garantizar entradas de datos de alta calidad y completas es crucial para la precisión de las valoraciones del AVM.

Por ejemplo, en países en desarrollo o regiones menos pobladas donde las transacciones inmobiliarias no se registran con tanta frecuencia, es posible que los AVMs no funcionen tan bien debido a la falta de datos completos. Esta limitación hace que sea difícil confiar únicamente en los AVMs en estas áreas.

Falta de análisis contextual

Los AVM pueden carecer de la capacidad para realizar un análisis contextual detallado que los tasadores humanos pueden proporcionar. Factores como renovaciones recientes, condiciones económicas locales, cambios en la dinámica del vecindario o proyectos de infraestructura próximos pueden no ser completamente capturados por los datos utilizados en los AVM. Por ejemplo, un AVM podría no tener en cuenta una renovación de alta calidad recientemente completada que mejora significativamente el valor de una propiedad. Los tasadores humanos, a través de su conocimiento local e inspecciones en el lugar, pueden ofrecer información sobre estos elementos contextuales, proporcionando una valoración más holística.

Por ejemplo, una propiedad ubicada cerca de una nueva escuela o desarrollo de transporte público podría ver aumentar su valor, pero un AVM podría no capturar completamente esta tendencia al alza sin los conocimientos contextuales que proporcionaría un tasador humano.

Sobredependencia en la tecnología

La creciente dependencia de los AVMs puede llevar a un papel disminuido para los tasadores humanos, lo que potencialmente reduce la profundidad de las evaluaciones de propiedades. Si bien los AVMs son herramientas valiosas, deberían complementar en lugar de reemplazar los métodos tradicionales de tasación. La sobredependencia en sistemas automatizados sin supervisión humana puede resultar en la omisión de matices y errores en las valoraciones. Por ejemplo, los AVMs podrían pasar por alto tendencias específicas del mercado local o características únicas de la propiedad que un tasador humano identificaría, como se mencionó anteriormente. La mejor práctica implica integrar los AVMs con las tasaciones tradicionales para combinar la eficiencia de la automatización con los conocimientos detallados de la experiencia humana.

Por ejemplo, durante el proceso de tasación de propiedades comerciales de alto valor, combinar los datos de los AVMs con una evaluación exhaustiva realizada por un tasador humano garantiza una valoración integral que tiene en cuenta tanto los aspectos cuantitativos como cualitativos de la propiedad.

Las consideraciones éticas de los AVMs y el Big Data

Si bien los Métodos de Valoración Automatizada y su uso de big data ofrecen numerosas ventajas en la valoración inmobiliaria, existen consideraciones éticas significativas que deben tenerse en cuenta.

Preocupaciones de privacidad

La ley española sobre la recopilación de datos y la privacidad personal – LEY ORGANICA:

Derecho a la intimidad ante la utilización de sistemas de geolocalización.

(art. 90).(BOE-A-2018-16673 Ley Orgánica 3/2018, de 5 de Diciembre, de Protección de Datos Personales y Garantía de Los Derechos Digitales., n.d.)

El uso de big data para alimentar los AVMs plantea importantes problemas de privacidad. Esto cuestiona la forma en que se recopilan, almacenan y utilizan los datos. En algunos casos, la información personal y/o sensible puede acumularse y luego explotarse sin el consentimiento de los propietarios de las propiedades. La ley española ha demostrado ser bastante restrictiva en lo que respecta a la privacidad de los datos personales. De hecho, la Ley Orgánica 7/2021 de protección de datos establece límites muy precisos para el uso y la recopilación de datos personales. Los AVM utilizados en España deben cumplir naturalmente con las leyes del país. Más adelante veremos que esto requiere vigilancia para evitar la acumulación de datos sensibles y su uso para evitar la discriminación y la categorización ilegal: raza, religión, orientación sexual, entre otros

Transparencia

La transparencia suele ser un tema cuestionable cuando se trata de AVMs. Más allá de la necesidad de proteger la propiedad intelectual, a menudo existe una falta de transparencia sobre qué datos se están utilizando y cómo se están procesando. Esta falta de transparencia a veces puede llevar al mal uso y la explotación de datos. En casos extremos, los AVM pueden recopilar y explotar datos que son ilegales o éticamente cuestionables. Veremos que esto puede afectar más tarde a la imparcialidad y sesgo del AVM.

Imparcialidad y Sesgo de los AVMs

Los AVMs que dependen de datos, si no están programados para analizar y procesar los datos de manera justa, pueden impulsar la perpetuación de sesgos y prácticas

discriminatorias existentes. La falta de transparencia y prácticas antidiscriminatorias claras pueden conducir exactamente a eso. Si los AVMs presentan un sesgo en la recopilación y procesamiento de datos, podemos esperar ver sesgos discriminatorios en las valoraciones; por ejemplo, las propiedades en vecindarios minoritarios pueden estar sujetas a una subvaloración repetida. Es importante recordar que los AVMs no consideran todos los factores en la evaluación de propiedades, sino solo los puntos de datos para los que han sido programados. Por lo tanto, si hay un sesgo preprogramado dentro del método de recopilación o procesamiento de datos, se puede esperar que haya un sesgo en los resultados de las valoraciones y una perpetuación adicional de dicho sesgo.

Desde un punto de vista ético, se espera que los proveedores de AVMs garanticen valoraciones de propiedades justas y equitativas en todas las comunidades y grupos socioeconómicos. La mitigación del sesgo es un factor clave que debería integrarse en la concepción de un AVM, no solo para garantizar valoraciones más precisas, sino también para asegurar la imparcialidad en la recopilación y procesamiento de datos.

Rendición de cuentas

Determinar la responsabilidad por los errores generados por las valoraciones de AVM puede ser complejo. Los errores en las valoraciones de AVM pueden tener un gran impacto en términos de pérdidas financieras o prácticas crediticias injustas. Por ejemplo, un AVM sesgado puede subvalorar o sobrevalorar una propiedad en una comunidad minoritaria, lo que lleva a tasas crediticias injustas o valoraciones injustas que afectan directamente a las personas. Es posible que alguien no pueda refinanciar una hipoteca a niveles justos debido a un error de AVM.

Una vez más, la transparencia se pone en duda, ya que puede hacer aún más difícil determinar la responsabilidad. La opacidad y la complejidad de los AVM pueden dificultar la comprensión del procesamiento de datos y, por lo tanto, determinar la responsabilidad por los errores de AVM

Impacto en la profesión de tasador

El impacto ético de los AVMs en los profesionales del sector es una preocupación importante que debe abordarse. Si bien los AVMs se consideran herramientas complementarias y no un reemplazo de los tasadores tradicionales, representan una amenaza para la profesión. El aumento en el uso de AVMs está provocando y seguirá provocando desplazamiento laboral y plantea cuestiones éticas sobre el futuro de la profesión.

La devaluación de habilidades también es una preocupación; la dependencia de los Modelos de Valoración Automática puede conducir a una devaluación gradual de la experiencia de los tasadores humanos. Esta devaluación puede provocar un deterioro en la calidad general de las valoraciones de propiedades. Aunque los AVMs están mejorando continuamente en su capacidad para procesar más datos y aumentar la precisión, la valoración de propiedades naturalmente implica un factor subjetivo. Aunque están cada vez más capacitados para reproducir los métodos de tasación humana, como lo demuestran los "AVMs de Emulación de Tasación", estos solo replican tales métodos de manera limitada. Por lo tanto, se puede cuestionar si la precisión y el comportamiento de los AVM pueden realmente replicar los de un tasador humano.

Regulación y cumplimiento normativo

Existen varias preocupaciones cuando se trata de la estandarización de los AVMs, su regulación y problemas de cumplimiento. Estos problemas pueden mitigarse mediante el cumplimiento de normas estándar, lo cual se logra creando AVMs que cumplan con las normas internacionales de valoración emitidas por el Consejo Internacional de Normas de Valoración (IVSC) o la Asociación Europea de Tasadores (TEGoVA).

La valoración conforme a las normas del IVS:

El IVSC ha establecido que, "al determinar si un AVM residencial puede cumplir con el IVS, primero es necesario tener en cuenta que la valoración es un proceso. Este proceso incluye varios requisitos en relación con el alcance del trabajo, investigaciones y cumplimiento, informes, bases de valor y enfoques y métodos de valoración, todos los cuales deben cumplirse para producir una valoración conforme al IVS".(Stokes, 2022)

El IVSC además establece que: "90.2. Al utilizar o crear un modelo de valoración, el tasador debe: (a) Mantener registros apropiados para respaldar la selección o creación del modelo, (b) Comprender y garantizar que el resultado del modelo de valoración, las suposiciones significativas y las condiciones limitantes sean consistentes con la base y el alcance de la valoración, y (c) Considerar los principales riesgos asociados con las suposiciones realizadas en el modelo de valoración. 90.3. Independientemente de la naturaleza del modelo de valoración, para cumplir con el IVS, el tasador debe garantizar que la valoración cumpla con todos los demás requisitos contenidos en el IVS."(Stokes, 2022).

La clave para garantizar el cumplimiento del AVM con el estándar IVS es considerarlo y aplicarlo desde su creación hasta su uso. El factor humano también es enfatizado por el IVSC, que destaca que el cumplimiento del IVS puede lograrse si el tasador utiliza el AVM como una herramienta y sigue el cumplimiento del IVS en todo momento, lo que a su vez produce una valoración conforme al IVS.

Una gran barrera que enfrentan los AVM para obtener el cumplimiento del IVS está relacionada con la entrada de datos. El estándar IVS dicta que, para que una valoración sea conforme al IVS, el tasador debe tener conocimiento y capacidad para controlar todas las entradas del modelo de valoración. Esto nos lleva de nuevo al problema de la transparencia, ya que la complejidad de la mayoría de los AVM, su falta de transparencia y la incapacidad de un individuo para controlar todas las entradas debido a la naturaleza de "Big Data" del AVM hacen extremadamente difícil que una valoración automatizada producida por un AVM sea conforme al IVS. En resumen, al considerar la posibilidad de que los AVM residenciales sean conformes al IVS, el Consejo Internacional de Normas de Valoración determinó que "es poco probable que un AVM residencial en aislamiento cumpla con estos requisitos de informes."(Stokes, 2022).

Conclusiones

Los Modelos de Valoración Automatizada (AVMs, por sus siglas en inglés) han transformado el campo de la valoración inmobiliaria mediante su capacidad para proporcionar evaluaciones rápidas, rentables y consistentes de propiedades. Este trabajo ha explorado varios tipos de AVMs, incluyendo Modelos Hedónicos, Modelos de Emulación de Tasación, Modelos de Índice, Modelos Mixtos y Modelos en Cascada, rastreando su evolución desde principios de los años 90 hasta la actualidad. Inicialmente limitados por datos y restricciones computacionales, los AVMs se han vuelto más sofisticados con los avances en la disponibilidad de datos y las tecnologías de aprendizaje automático.

El estudio comparó específicamente diferentes metodologías de AVMs como la Regresión de Vectores de Soporte (SVR), Bosques Aleatorios (RF), Impulso Extremo de Gradiente (XGBoost) y Redes Neuronales Profundas (DNN). Entre estos, el modelo XGBoost destacó por su precisión superior y eficiencia computacional, resaltando las ventajas de los algoritmos avanzados de aprendizaje automático en la valoración de propiedades.

Se probaron dos modelos prácticos utilizando datos del Condado de King, Washington, EE. UU. El Modelo 1, utilizando Regresión Lineal Múltiple (MLR) con 500 observaciones de entrenamiento y 100 de prueba, explicó el 57.23% de la variabilidad en los precios de las casas ($R^2 = 0.5723$). Los predictores significativos incluyeron dormitorios, baños, pies cuadrados de espacio habitable, grado y año de construcción, aunque el modelo enfrentó limitaciones debido a la multicolinealidad y variables no significativas como el tamaño del lote y la condición. El Modelo 2 mejoró esto utilizando 21,611 observaciones, logrando un R^2 de 0.653 y explicando el 65.3% de la variabilidad en los precios de las casas. Los predictores significativos para este modelo incluyeron factores adicionales como pisos, presencia de frente de agua y vista. A pesar de un mejor rendimiento, este modelo tuvo un porcentaje de error promedio del 31%, una desviación estándar del 45% y una varianza del 20%, indicando espacio para una mejora adicional.

Las ventajas de los AVMs son evidentes en su capacidad para mejorar la eficiencia de las transacciones, reducir los costos de valoración y minimizar el sesgo humano a través del procesamiento automatizado de datos. Sin embargo, los AVMs también enfrentan desafíos, especialmente con propiedades que tienen características únicas o se encuentran en áreas con datos limitados. Su fiabilidad depende en gran medida de la calidad y completitud de los datos, y es posible que no capturen completamente factores contextuales como renovaciones recientes o condiciones económicas locales. La confianza exclusiva en los AVMs puede pasar por alto matices que los tasadores humanos podrían captar.

Además, el uso de AVMs plantea consideraciones éticas significativas, incluyendo preocupaciones de privacidad, transparencia, imparcialidad, sesgo y responsabilidad. Hay que asegurar que los AVMs sean transparentes y justos es crucial para prevenir prácticas discriminatorias y garantizar valoraciones de propiedades equitativas en diferentes comunidades y grupos socioeconómicos.

En conclusión, aunque los AVMs han avanzado significativamente en la valoración inmobiliaria al proporcionar soluciones eficientes y rentables, deben ser continuamente refinados y gestionados éticamente para garantizar valoraciones precisas y justas. La integración de los AVMs con métodos de valoración tradicionales puede aprovechar las fortalezas de ambos enfoques, proporcionando una herramienta más completa para la valoración de propiedades. Esta integración garantizará que los avances tecnológicos mejoren la precisión y confiabilidad de las valoraciones de propiedades, abordando tanto las limitaciones actuales como las preocupaciones éticas.

Bibliografía

Babb, C. H. (n.d.). AVMs: Then, Now, and in the Future. *DSNews*. Retrieved 24 May 2024,

from <https://dsnews.com/featured/08-09-2021/avms-then-now-and-in-the-future>

BOE-A-2018-16673 Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos

Personales y garantía de los derechos digitales. (n.d.). Retrieved 25 May 2024,

from <https://www.boe.es/buscar/act.php?id=BOE-A-2018-16673>

Carranza, J. P., Piumetto, M. A., Lucca, C. M., & Da Silva, E. (2022). Mass appraisal as

affordable public policy: Open data and machine learning for mapping urban land values. *Land Use Policy*, 119, 106211.

<https://doi.org/10.1016/j.landusepol.2022.106211>

Dettbarn, Davis, L. Escobar, Finnsson, Ripperger, Clark. (2003). *AVM Standard*.

https://www.iaao.org/wcm/About/News/wcm/About_Us_Content/News/AVM_standard.aspx

DODD-FRANK WALL STREET REFORM AND CONSUMER PROTECTION ACT. (n.d.).

Retrieved 24 May 2024, from

<https://www.congress.gov/111/plaws/publ203/PLAW-111publ203.pdf>

EAA_Standards_3rd_Edition.pdf. (n.d.). Retrieved 24 May 2024, from

https://www.europeanavmalliance.org/files/eaa/Downloads/EAA_Standards_3rd_Edition.pdf

Garbin, C., Zhu, X., & Marques, O. (2020). Dropout vs. batch normalization: An empirical

study of their impact to deep learning. *Multimedia Tools and Applications*,

79(19), 12777–12815. <https://doi.org/10.1007/s11042-019-08453-9>

Jafary, P., Shojaei, D., Rajabifard, A., & Ngo, T. (2024). Automated land valuation models:

A comparative study of four machine learning and deep learning methods based on a comprehensive range of influential factors. *Cities*, 151, 105115.

<https://doi.org/10.1016/j.cities.2024.105115>

Kc_house_data. (n.d.). Retrieved 4 June 2024, from

<https://www.kaggle.com/datasets/shivachandel/kc-house-data/data>

Rep. Frank, B. [D-M.-4. (2010, July 21). *Text - H.R.4173 - 111th Congress (2009-2010):*

Dodd-Frank Wall Street Reform and Consumer Protection Act (2009-12-02)

[Legislation]. <https://www.congress.gov/bill/111th-congress/house-bill/4173/text>

Standard_on_Automated_Valuation_Models.pdf. (n.d.). Retrieved 30 May 2024, from

https://www.iaao.org/media/standards/Standard_on_Automated_Valuation_Models.pdf

Statistical Valuation Standards—European AVM Alliance. (n.d.). Retrieved 24 May 2024,

from <https://www.europeanavmalliance.org/en/avm-standards.html>

Stokes, R. (2022, November 30). Perspectives Paper: Automated Valuation Models and

Residential Valuations. *International Valuation Standards Council*.

<https://www.ivsc.org/perspectives-paper-automated-valuation-models-and-residential-valuations/>

The State of Automated Valuation Models in the Age of Big Data. (n.d.). Retrieved 22 May

2024, from <https://www.mba.org/docs/default-source/uploadedfiles/member-white-papers/stateofautomatedvaluationmodels-final>

Tringham, M. (2007, March 23). Hard-pressed discover a 15-minute solution. *Financial*

Times. <https://www.ft.com/content/b12d86f8-d936-11db-9b4a-000b5df10621>

Vapnik, V. (2013). *The Nature of Statistical Learning Theory*. Springer Science & Business Media.

ANNEXO EXCEL

<file:///C:/Users/natha/OneDrive/Desktop/TFG%20Model%20V4.xlsx>

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Nathan Duchene, estudiante de E4 de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado , declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. **Brainstorming de ideas de investigación:** Utilizado para idear y esbozar posibles áreas de investigación.
2. **Metodólogo:** Para descubrir métodos aplicables a problemas específicos de investigación.
3. **Estudios multidisciplinares:** Para comprender perspectivas de otras comunidades sobre temas de naturaleza multidisciplinar.
4. **Generador previo de diagramas de flujo y contenido:** Para esbozar diagramas iniciales.
5. **Revisor:** Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
6. **Traductor:** Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 05.06.24

Firma: Nathan Duchene