Predicción de precio de una casa en base a sus características

Diciembre 2023

CODERHOUSE – DATA SCIENCE

COMISION: 42390

AUTOR: GABRIELA MONTES HERNANDEZ

Tabla de Contenido

Abstract

For real estate agencies, selling a house can become a complicated task, **from accepting the initial conditions to setting a price**. Even more important if the price of the house is well above the market, making the process more difficult and late. In fact, according to studies conducted in various sectors worldwide, one of **the most common mistakes** made by real estate agents is to set **a price higher than it should be**, thinking that they will get a higher commission even if they manage to negotiate prices. However, **setting a price below the market will not guarantee a sale.**

Therefore, for this research, variables will be analyzed with special emphasis on the type and style of housing and conditions.

Likewise, it is established that this research is directed to the general public that may not **have a deep technical knowledge in real estate sales but a technical level in data analysis**, which will allow students or people interested in this field, to understand the concepts of the market.

Hipótesis

Como se ha mencionado anteriormente, fijar el precio correcto de una vivienda, permite a las inmobiliarias a **obtener mejor posicionamiento en el mercado** debido a las mejores ofertas que estas puedan otorgar, aun mas basándose en las características que estas puedan tener. Es por ello que surgen las siguientes preguntas:

1. ¿Las condiciones físicas de una vivienda puede **influir en la existencia** de un comprador?

2. ¿Es posible que, al mejorar el precio de una vivienda, **disminuya los escases de compradores**, permitiendo incluso que la clase media tenga acceso a adquirir una?

3. ¿La zona geográfica de las viviendas es un **factor determinante para su precio**?

Introducción

Contexto Comercial: En el mundo actual, la toma de decisiones informada y respaldada por datos precisos es esencial en el mercado. En este caso, **basándonos en el mercado inmobiliario**, ya sea que se esté buscando comprar una casa como inversión o como un hogar, **contar con información sólida es crucial**. Es aquí donde nuestro enfoque centrado en los datos y la analítica se centra en una amplia gama de propiedades, incluyendo detalles como la ubicación, el tamaño del lote, la cantidad de habitaciones y baños, la edad de la construcción, las mejoras realizadas y muchas otras características.

Problema Comercial: Cuando una casa se valora incorrectamente debido a una falta de datos precisos sobre sus características, **puede llevar a la sobrevaloración**, lo que alejará a los posibles compradores, o a la subvaloración, resultando una pérdida de ingresos para el vendedor.

Contexto Analítico: En el pasado, la valoración de las propiedades solía depender en gran medida de la intuición de los expertos y de comparaciones subjetivas. Sin embargo, en la actualidad, **la analítica de datos se ha convertido en el pilar fundamental** para determinar el precio de las casas de manera precisa y objetiva. Estos factores incluyen el tamaño del lote, la superficie habitable, el número de habitaciones y baños, la edad y estado de la construcción, las mejoras realizadas, la proximidad a servicios y escuelas, y otros atributos específicos. Cada uno de estos elementos es **evaluado de manera individual** y en combinación para determinar su impacto en el precio final.

Definición de Objetivos

Para el presente proyecto, se tiene establecido los siguientes objetivos:

1. **Lograr** una precisión significativa en la estimación del precio de una casa en función de sus características.
2. **Asegurar** que el modelo incluya todas las características relevantes que influyen en el precio de una casa.
3. **Desarrollar** el modelo para que sea capaz de hacer predicciones tanto a corto plazo como a largo plazo.

Solución Propuesta

Se propone **crear un modelo de regresión lineal** predictivo, sólido y útil para evaluar el precio de las casas en función de sus características.

Para ello se usará datos como: clasificación general de zonificación, pies lineales de calle conectados a la propiedad, tamaño del lote en pies cuadrados, tipo de camino y callejón de acceso a la propiedad, forma general y planitud, entre otras variables importantes.

Modelado

Data Acquisition

La adquisición de datos para este proyecto se obtuvo mediante la recopilación de una fuente común de datos, como de fuentes públicas, registros de propiedades y bases de datos gubernamentales. Adicional, las agencias inmobiliarias suelen mantener registros detallados sobre propiedades que están en venta o que han sido vendidas en el pasado. Los datos se obtuvieron de: https://www.kaggle.com/competitions/home-data-for-ml-course/data?select=train.csv

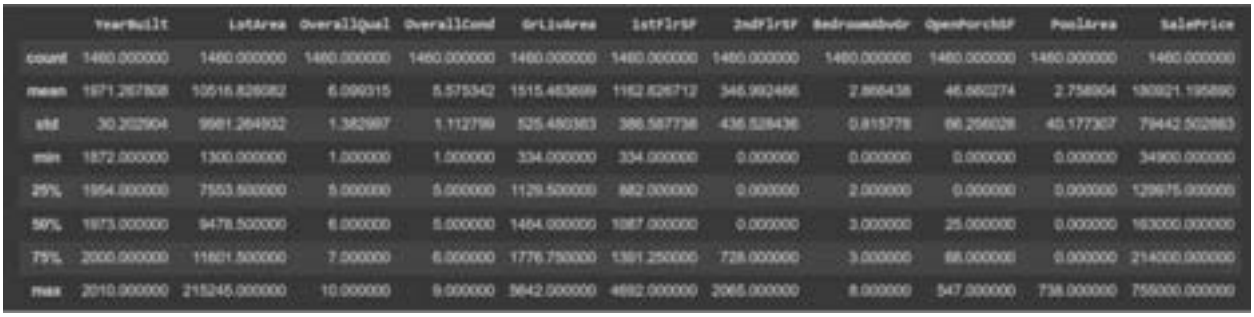
Desglosado en:

- train.csv: el conjunto de entrenamiento.
- test.csv: el conjunto de prueba.
- data_description.txt: descripción completa de cada columna, preparada originalmente por Dean De Cock pero ligeramente editada para que coincida con los nombres de las columnas utilizadas aquí.
- sample_submission.csv: un envío de referencia a partir de una regresión lineal sobre el año y mes de venta, los pies cuadrados del lote y el número de habitaciones.

Estadística Descriptiva

Contiene 1460 filas y 81 columnas. En este último dataset, se encuentra la variable target (SalePrice).

El dataset train contiene **3 variables float64, 35 variables int64 y 43 variables object.**



	YearBuilt	LotArea	OverallQual	OverallCond	GrLivArea	1stFlrSF	2ndFlrSF	BedroomAbvGr	OpenPorchSF	PoolArea	SalePrice
count	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000	1460.000000
mean	1971.267808	10516.828082	6.099315	5.575342	1515.463699	1162.626712	346.992466	2.866438	46.660274	2.758904	180921.195890
std	30.202904	9981.264932	1.382997	1.112799	525.480383	386.587738	436.528436	0.815778	66.256028	40.177307	79442.502883
min	1872.000000	1300.000000	1.000000	1.000000	334.000000	334.000000	0.000000	0.000000	0.000000	0.000000	34900.000000
25%	1954.000000	7553.500000	5.000000	5.000000	1129.500000	882.000000	0.000000	2.000000	0.000000	0.000000	129975.000000
50%	1973.000000	9478.500000	6.000000	5.000000	1464.000000	1087.000000	0.000000	3.000000	25.000000	0.000000	163000.000000
75%	2000.000000	11601.500000	7.000000	6.000000	1776.750000	1391.250000	728.000000	3.000000	68.000000	0.000000	214000.000000
max	2010.000000	215245.000000	10.000000	9.000000	5642.000000	4692.000000	2065.000000	8.000000	547.000000	738.000000	755000.000000

Figura 1. Estadística descriptiva básica de las variables importantes

Figura 1

- **LotArea (Área del lote):** muestra que el área de los lotes varía significativamente, desde 1.300 hasta 215.245 pies cuadrados, con una mediana en 9.478,5 pies cuadrados.
- **OverallQual (Calidad general):** indica que la calidad general de las propiedades tiende a estar en el rango de 1 a 10, con una mediana de 6.
- **OverallCond (Condición general):** muestra que la condición general de las propiedades tiende a estar en el rango de 1 a 9, con una mediana de 5.
- **GrLivArea (Área habitable sobre el nivel del suelo):** El área habitable tiende a variar desde 334 hasta 5.642 pies cuadrados, con una mediana de 1.464 pies cuadrados.

Data Wrangling y EDA

Se realiza una verificación de correlación entre variables, a través de un gráfico de calor:

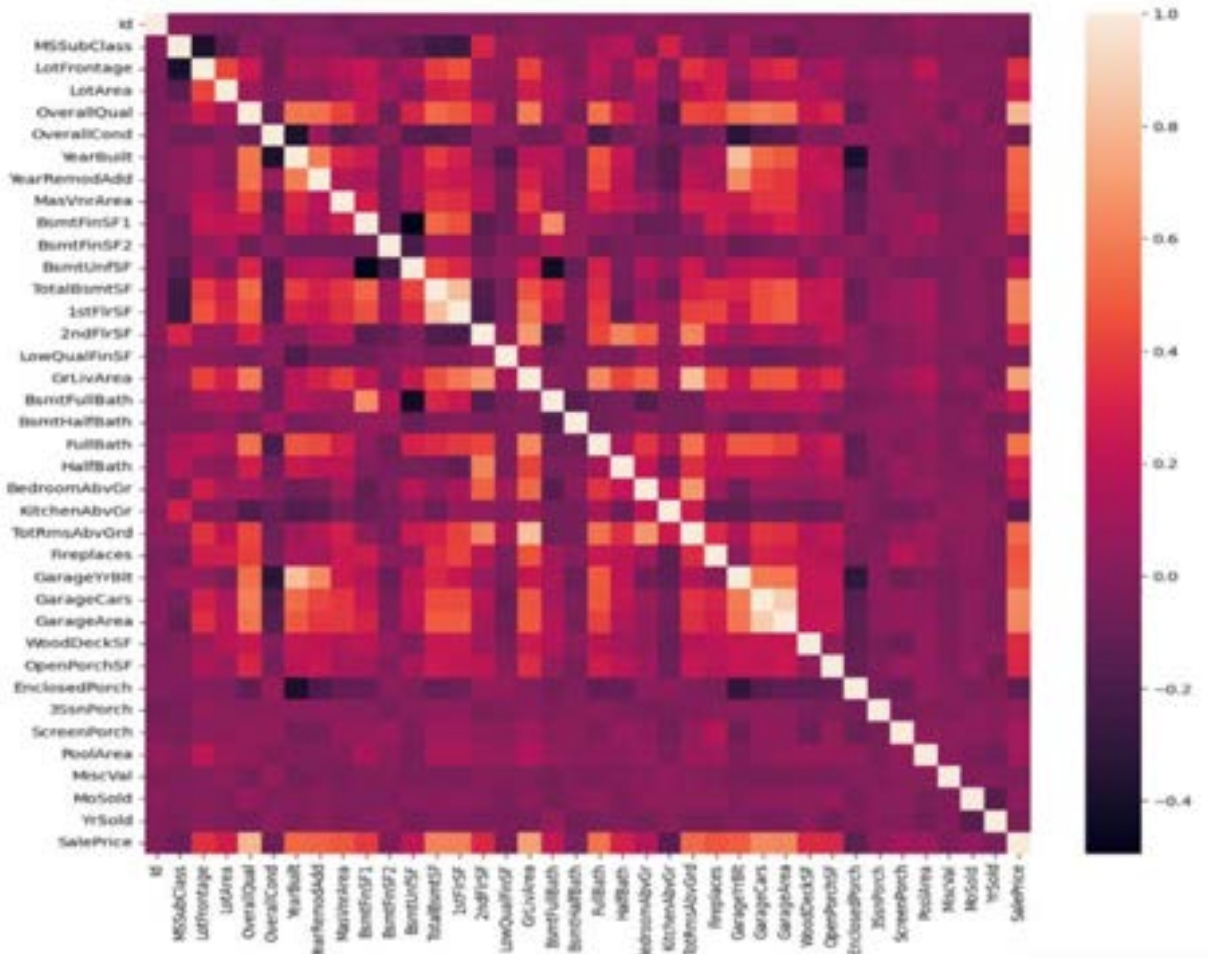


Gráfico 1. Gráfico de calor

Gráfico 1

Con el grafico de calor, no solo se observa las variables de alta correlación, sino también se puede visualizar aquellas con baja correlación y así se poder eliminarlas usando la formulación "drop".

Se aplica la formulación "drop" tanto para el dataset de entrenamiento como para el dataset de testeo. Eliminando las siguientes variables:

['Id','MSSubClass','OverallCond','BsmtFinSF2','LowQualFinSF','BsmtHalfBath','3Ssn Porch','ScreenPorch','PoolArea','MiscVal','MoSold','YrSold']

Se procede a **buscar solo los valores numéricos** a través de la función "select_dtypes" y se calcula la media para reemplazar los valores Nan a través de la función "mean()" y "fillna (medias_)".

Se evalúan los valores atípicos considerados mediante el gráfico de calor, a través de gráficos como: scatterplot, boxplot y barplot, comparado con la variable SalePrice (Precio de venta):

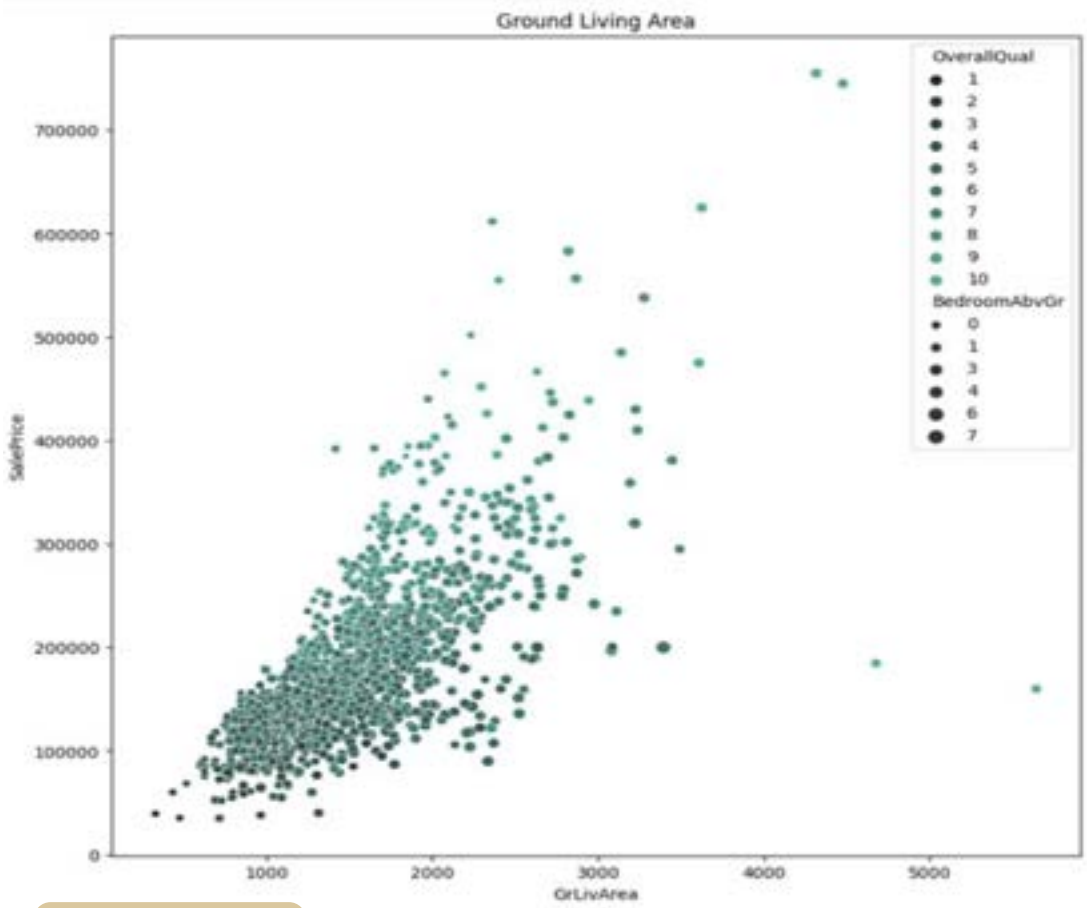


Gráfico 2. GrLivArea vs SalePrice

Gráfico 2

Muestra una concentración de puntos por **debajo de los 400.000**, lo que indica que la mayoría de las casas tienen superficie habitable por debajo de este valor, es decir, **son accesibles**.

Se puede concluir que los **valores por encima de 400.000** considerados valores atípicos, posiblemente se traten de casas con superficies más grandes, lo que tiene sentido en establecer que a **mayor espacio** o mientras más grande la casa, **mas es su valor**. También se puede apreciar que, en los puntos con mayor concentración, tienden a reflejar calificaciones bajas, por ejemplo, **mientras más bajo el precio** de 100.000 las **calificaciones generales oscilan entre 1 a 5**.

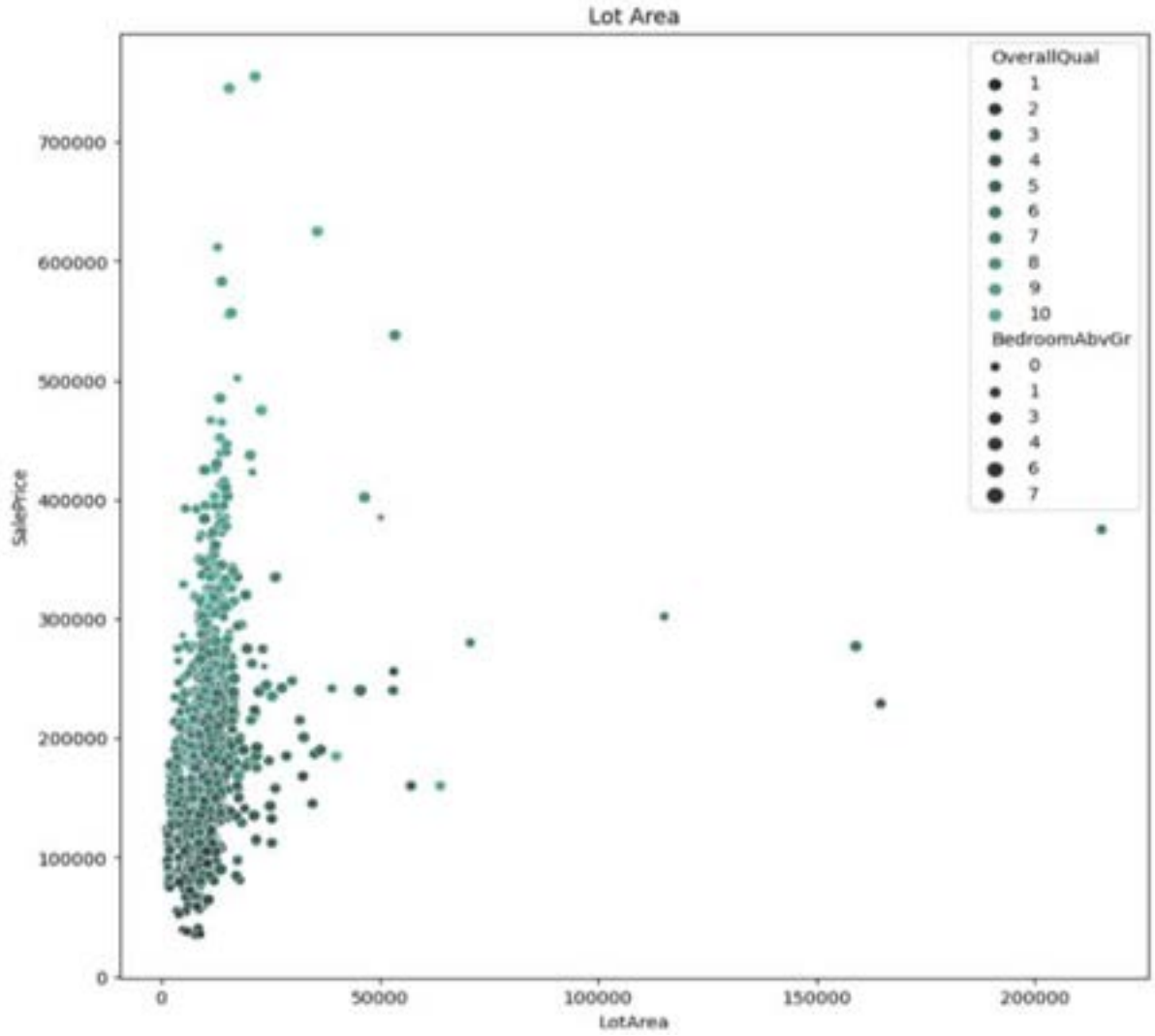


Gráfico 3. LotArea vs SalePrice

Gráfico 3

El **gráfico 3** nos muestra una concentración de **puntos por debajo de 600.000 y un área de 50.000**, lo que indica que la mayoría de las casas tienen un lote de área habitable por debajo de este valor, es decir, es accesible una casa con un lote menor pero grandes características.

Según la **concentración de puntos**, tienden a dispersarse hacia arriba y hacia la derecha, demostrando que no varía tanto el lote de área indiferente del precio, se puede cuestionar que hay otros factores que generan precios elevados, como: cantidad de habitaciones, baños, estacionamientos, ubicación geográfica, entre otros.

También se puede apreciar que los puntos alejados de la concentración, tienen la **mejor evaluación entre 7 a 10**.

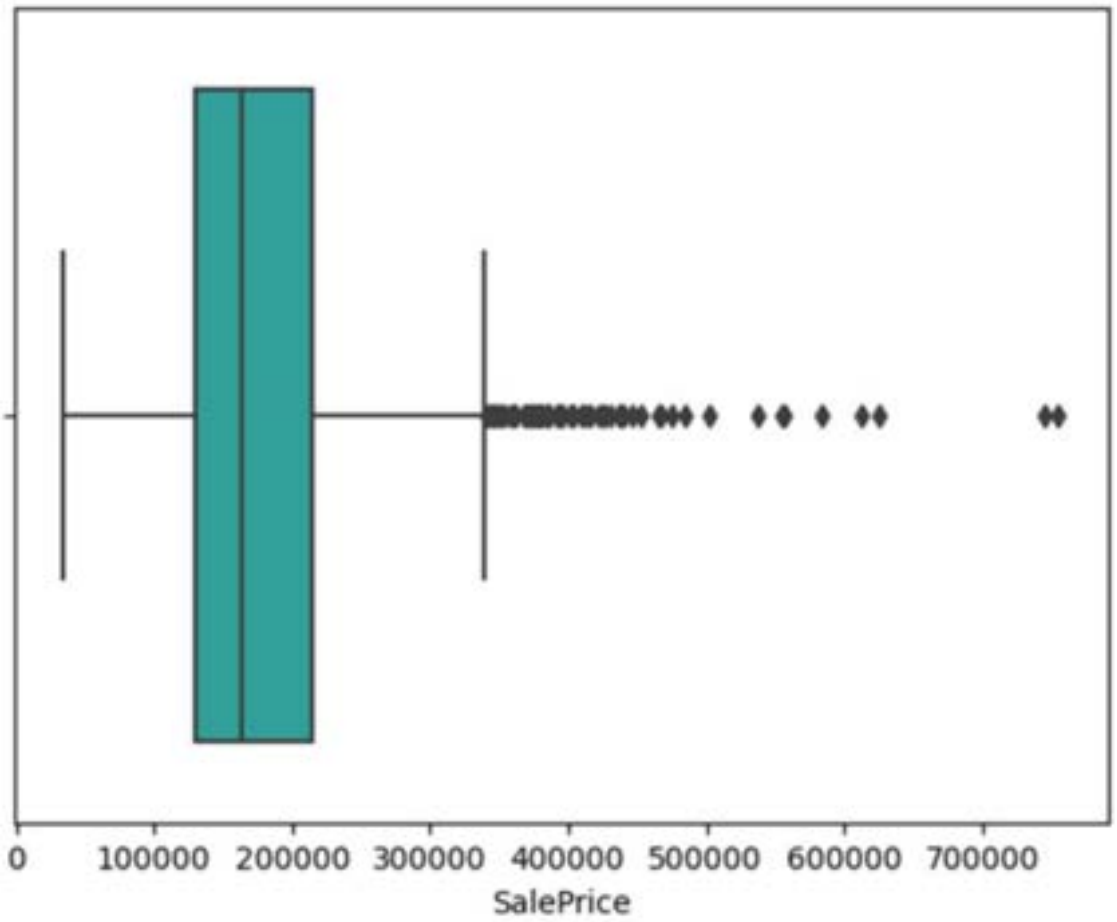


Gráfico 4. Rango de SalePrice

Gráfico 4

El **gráfico 4** de caja representa un intercuartílico de rango **entre 50.000 a 350.000** aproximadamente, donde **la mediana representa un valor** de 150.000 y los bigotes de extienden hasta un precio de 350.000, demostrando que la concentración de **puntos afuera de los bigotes** contiene valore atípicos por sobre los 450.000 pudiendo afectar la precisión de las predicciones.

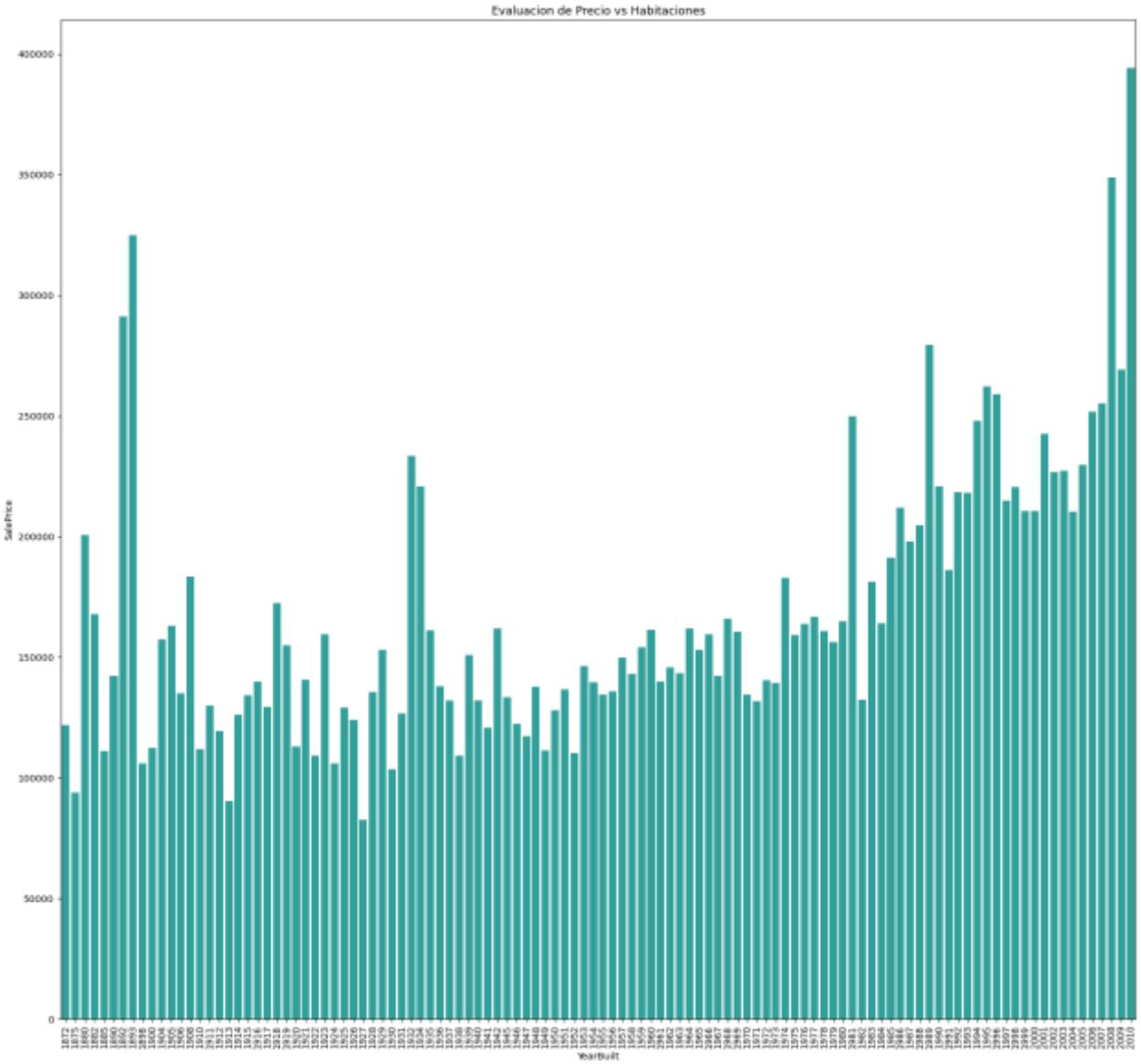


Gráfico 5. YearBuilt vs SalePrice

Gráfico 5

El **gráfico de barras (5)** nos muestra que el año con menor precio de venta fue en 1927. Realizando una evaluación general, se puede denotar que durante **102 años los precios se habían mantenido** con un promedio de 200.000 a 260.000 en su valor, fue solo hasta **el 2008 y 2010 donde los precios se elevaron** a 350.000 en adelante. Con esto, se podrá determinar que posiblemente el factor económico transcurrido durante esos años de estudio (1898 a 1991), fue en mejor medida accesible para el usuario tener una casa en relación a precio - calidad.

Menor precio: año 1927

Mayor precio: año 2010

Entrenamiento del Modelo (Machine Learning)

Regresión Lineal

Como primera iteración se procede a evaluar el modelo mediante "Regresión lineal". Realizando la búsqueda de hiperparámetros, el mejor polinomio y buscando los resultados a través de Cross-Validation.

Se debe eliminar la columna faltante en el dataset de testeo porque al momento de realizar la predicción, genera error mencionando específicamente que la característica "SalePrice" está ausente en los datos de evaluación o predicción, pero estaba presente durante el entrenamiento.

Se considera varios valores para el parámetro 'degree'. En este caso, los grados 1, 2, 3 y 4 son los evaluados. Como resultado, en búsqueda del mejor polinomio, se obtiene:

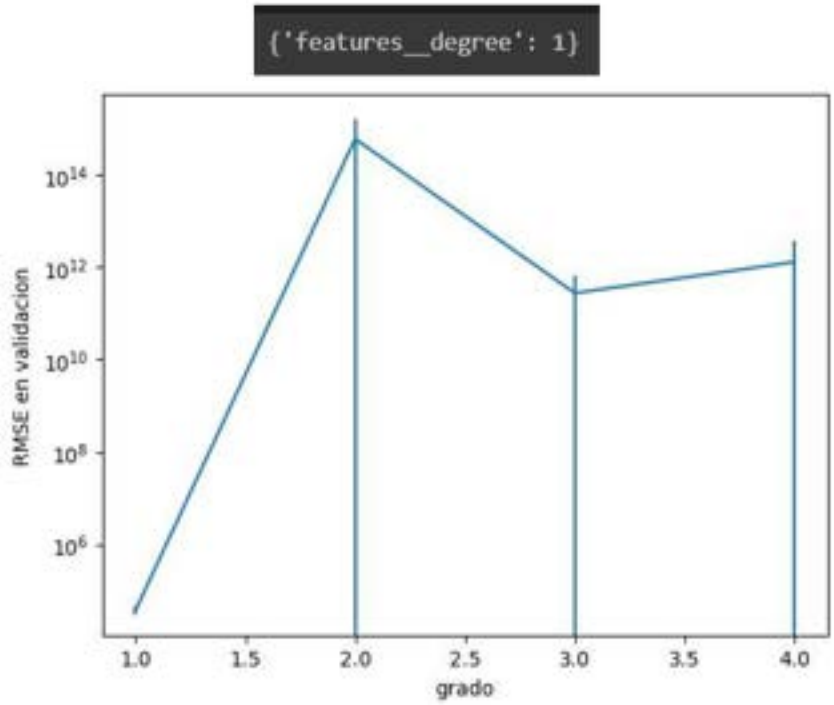


Gráfico 6. Resultados del CV

Gráfico 6

En el **gráfico 6** se puede observar que el **grado 1**, es el parámetro en donde **el polinomio se hace más pequeño**. Igualmente, se refleja que, si se cambia a otros polinomios de grados mayores, comienza una **tendencia de error.**

Menor precio: año 1927

Mayor precio: año 2010

Al entrenar el modelo con el mejor estimador, arroja los siguientes resultados:

	TRAIN	TEST
Raíz del error cuadratico medio (RMSE)	5.724103280166372e-13	5.731653799103435e-13
Error Absoluto Medio MAE	4.369728962234065e-13	4.433045949212503e-13
Error Cuadrático Medio MSE	4.385657645421329e-24	4.476233921534658e-24
Coeficiente de determinación R2	1.0	1.0

Regresión Lineal

Los valores generados **son muy pequeños**, significando que, tanto el error absoluto, error cuadrático y la raíz cuadrada promedio entre los valores predichos y los valores reales **son prácticamente cero**.

📌Este **MAE y MSE** tan pequeño sugiere que el modelo es preciso y que las predicciones son idénticas a los valores reales en el conjunto de datos de prueba.

📌El valor **r2 y r3 = 1.0**, indica que el modelo de regresión se ajusta perfectamente a los datos, es decir, puede explicar todas las variaciones en los datos de prueba.

Randomized Search CVA

Como segunda iteración se procede a evaluar el modelo mediante "Randomized Search CVA". Realizando la búsqueda de hiperparámetros a través de árbol de decisión y luego una búsqueda aleatoria con **Validación Cruzada.**

Se busca la mejor profundidad máxima con "max_depth" lo que permite controlar la profundidad máxima permitida para cada árbol en el modelo. Probando con diferentes valores **(3, 5, 10, 12, 15).**

Se busca controlar el número mínimo de muestras necesarias para dividir un nodo interno con "min_samples_split". Se prueba con diferentes valores para encontrar el adecuado **(2, 5, 10, 15, 20).**

Po último se establece el número mínimo de muestras requeridas para estar en un nodo hoja con "min_samples_leaf". Se prueba con diferentes valores para encontrar el adecuado **(1, 2, 4, 6, 8, 10).**

Como resultado, se obtiene lo siguiente:

Validación Cruzada

regression__min_samples_split = 10, significa que un nodo solo se dividirá si hay al menos 10 muestras en él.

regression__min_samples_leaf = 4, significa que cada hoja debe tener al menos 4 muestras.

regression__max_depth = 15, indica que el árbol tiene una profundidad máxima de 15 niveles. Esto limita la cantidad de divisiones y nodos en el árbol.

Al entrenar el modelo con el mejor estimador, arroja los siguientes resultados:

	TRAIN	TEST
Raíz del error cuadratico medio (RMSE)	149.09486522814484	113.11088151838194
Error Absoluto Medio MAE	55.33948095481147	68.83255191886546
Error Cuadrático Medio MSE	167915.26307685405	33711.95963629017
Coeficiente de determinación R2	0.5451109033354489	0.3324821320431422

Randomized Search CVA

Los valores generados **son muy grandes**, significando que, tanto el error absoluto, error cuadrático y la raíz cuadrada promedio entre los valores predichos y los valores reales **no se encuentran ajustados adecuadamente.**

📌Este **MAE y MSE** tan grande sugiere que el modelo no es preciso y que las predicciones **no son idénticas a los valores reales** en el conjunto de datos de prueba.

📌Los valores R2, indican que el rendimiento del modelo es deficiente, posiblemente por datos no vistos.

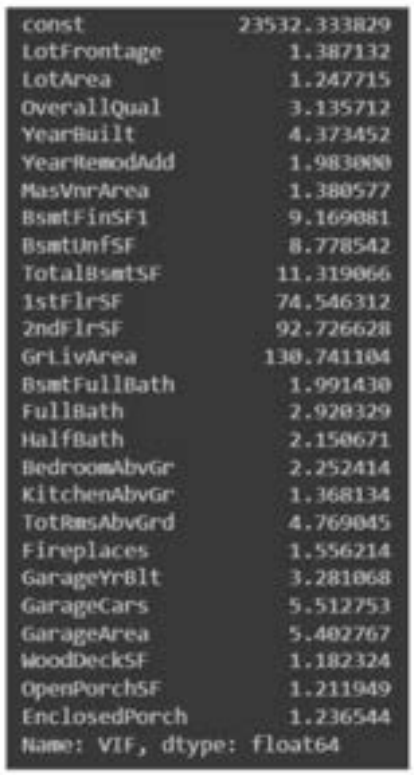
Regresión Lineal Múltiple

Como tercera iteración se procede a evaluar el modelo mediante "Regresión Lineal Múltiple".

Se evalúa a través del Ordinary Least Squares, es decir se busca ajustar la línea a un conjunto de datos para visualizar las diferencias entre los datos reales y las predicciones del modelo.

Adicional, se usa la función add_constant para ajustar el modelo y que este tenga un término independiente.

Como resultado se obtiene lo siguiente:



```
const            23532.333829
LotFrontage          1.387132
LotArea              1.247715
OverallQual          3.135712
YearBuilt            4.373452
YearRemodAdd         1.983000
MasVnrArea           1.380577
BsmtFinSF1           9.169081
BsmtUnfSF            8.778542
TotalBsmtSF         11.319066
1stFlrSF            74.546312
2ndFlrSF            92.726628
GrLivArea          130.741104
BsmtFullBath         1.991430
FullBath             2.920329
HalfBath             2.150671
BedroomAbvGr         2.252414
KitchenAbvGr         1.368134
TotRmsAbvGrd         4.769045
Fireplaces           1.556214
GarageYrBlt          3.281068
GarageCars           5.512753
GarageArea           5.402767
WoodDeckSF           1.182324
OpenPorchSF          1.211949
EnclosedPorch        1.236544
Name: VIF, dtype: float64
```

Figura 3. Ajuste del modelo

Figura 3

Se evidencia que hay tres variables con un VIF muy elevado: 1stFlrSF , 2ndFlrSF , GrLivArea.

Esto significa que la colinealidad es muy fuerte. Indica que las variables están altamente correlacionadas con otras variables y puede afectar significativamente las estimaciones de los coeficientes de regresión.

Al entrenar el modelo con el mejor estimador, arroja los siguientes resultados:

	TRAIN	TEST
Raíz del error cuadratico medio (RMSE)	35585.695172642205	35318.71030423272
Error Absoluto Medio MAE	21908.75311483857	22690.13822373087
Error Cuadrático Medio MSE	1266341700.9202106	1247411297.5543149
Coeficiente de determinación R2	0.7914266416525908	0.8219334381631942

Regresión Lineal Múltiple

Se evidencia que los valores no mejoraron respecto a los modelos estudiados anteriormente.

- **Para el RMSE, ambos conjuntos tienen valores similares, lo que indica una consistencia en el rendimiento del modelo.**
- **Para el MAE, ambos conjuntos tienen valores similares, y en ambos casos, el MAE es más bajo que el RMSE.**
- **Para los valores del MSE, dan más peso a los errores grandes que el MAE.**
- **Para ambos conjuntos del R2, tienen valores bastante altos, indicando que una gran proporción de la variabilidad en la variable dependiente está siendo explicada por el modelo.**

Conclusiones Generales

La aplicación de diferentes técnicas de regresión, incluyendo **regresión lineal, Randomized Search CV, y regresión lineal múltiple**, ha proporcionado una serie de métricas evaluativas para cada modelo. Los resultados obtenidos en las métricas de error (RMSE y MAE) y en la capacidad predictiva (R2) ofrecen una **visión integral del rendimiento** de cada modelo en el conjunto de datos utilizado.

Observando los valores de RMSE, se destaca que el **Modelo 1 exhibe un error cuadrático medio** extremadamente bajo, indicando una ajustada capacidad del modelo para predecir los valores target. En contraste, el **Modelo 3 presenta un RMSE** significativamente más elevado, lo que sugiere una mayor discrepancia entre las predicciones y los valores reales.

La métrica MAE confirma esta tendencia, donde el **Modelo 1** y el **Modelo 2 exhiben valores cercanos a cero**, indicando una menor magnitud de error absoluto medio en comparación con el Modelo 3.

En cuanto al **coeficiente de determinación (R2)**, que mide la proporción de la variabilidad en la variable dependiente explicada por el modelo, se observa que todos los modelos presentan un rendimiento aceptable. Sin embargo, se aprecia una clara **diferencia entre el Modelo 1** con un **R2 perfecto de 1.0** y el **Modelo 3 con un R2 ligeramente menor**, indicando una explicación ligeramente menor de la variabilidad en el conjunto de datos.

En resumen, los resultados sugieren que el **Modelo 1 ha logrado un ajuste** excepcionalmente bueno a los datos, minimizando tanto el error absoluto como el cuadrático medio. Mientras que el **Modelo 2 presenta un rendimiento intermedio**, el **Modelo 3 muestra un rendimiento inferior** en términos de precisión predictiva. Estos hallazgos destacan la importancia de seleccionar cuidadosamente la técnica de regresión y el proceso de ajuste del modelo para lograr resultados óptimos en la predicción de la variable dependiente.

Insights & Recomendaciones

Insights Precios

El análisis de precios de casas ofrece valiosas perspectivas para compradores, vendedores e inversores inmobiliarios, por ejemplo:

✓ **La ubicación** es fundamental ya que pueden tener precios muy distintos. Varían significativamente según la ubicación de barrios, ciudades e incluso vecindarios específicos.

✓ **La tendencia estacional** puede afectar los precios. Las ventas de viviendas tienden a aumentar en primavera y verano, lo que puede llevar a precios más altos durante esos meses.

✓ Las casas tienden a aumentar de valor con el tiempo así que **los inversores pueden aprovechar** esto para obtener ganancias a largo plazo.

Insights Características

✓ **Las características** de una casa, como el tamaño, el número de habitaciones, los acabados y las comodidades, afectan significativamente su precio. Cuanto más grande o más lujosa sea una propiedad, generalmente mayor será su precio de venta.

✓ **Las renovaciones y mejoras,** como una cocina actualizada o un baño renovado, pueden aumentar el valor de una casa.

✓ **Las características personalizadas**, como piscinas, paisajismo, cuidado y características únicas de diseño, pueden aumentar el precio, pero su valor puede varias según el gusto personal de los compradores.

✓ **Las tendencias del mercado inmobiliario** local también pueden afectar como se valoran las características, por ejemplo, en un mercado competitivo, las características de calidad pueden ser más valoradas.

Recomendaciones

El análisis de precio vs calidad es un proceso complejo que implica múltiples variables y consideraciones. La calidad de una casa puede ser subjetiva en algunos aspectos, pero con una metodología sólida y datos confiables, se puede obtener valiosos insights sobre como la calidad influye en el precio de las casas.

Por ejemplo:

✓ Tener datos precisos y confiables.

✓ Considerar utilizar métricas objetivas siempre que sea posible.

✓ Tener en cuenta el contexto y las expectativas de los compradores en el área en cuestión.

✓ Investigar a fondo los productos o servicios que se está considerando.

✓ Comprender las características, las especificaciones técnicas y las revisiones de otros consumidores. Esto ayudara a determinar la calidad intrínseca de lo que se está adquiriendo.

Para el análisis de modelado se puede verificar la capacidad del modelo y así generalizar en situaciones del mundo real, en este caso, se puede considerar otras métricas de evaluación y técnicas de validación cruzada para obtener una imagen más completa del rendimiento del modelo.