UNIVERSIDAD CATÓLICA BOLIVIANA "SAN PABLO"
UNIDAD ACADÉMICA REGIONAL LA PAZ
DEPARTAMENTO DE POSGRADO
MAESTRÍA EN CIENCIA DE DATOS

PROYECTO 1 - CASO DE ESTUDIO

"Predicción de Precios en Bienes Raíces - Mercado Australiano - Regresión Avanzada"

MÓDULO de Machine Learning

Docente: Jheser Guzman Ph.D.

Maestrantes: Virginia Mercedes Fernández Daza
Marco Antonio Velásquez Rocha
Ivan Israel Machicado Quiroga

noviembre, 2024

Índice

1. Antecedentes.

Una empresa de vivienda con sede en EE. UU. llamada Surprise Housing ha decidido ingresar al mercado australiano. La empresa utiliza el análisis de datos para comprar casas a un precio inferior a sus valores reales y venderlas a un precio más alto. Con el mismo propósito, la empresa ha recopilado un conjunto de datos de la venta de casas en Australia.

La compañía está buscando posibles propiedades para comprar e ingresar al mercado. Debe construir un modelo de regresión utilizando la regularización para predecir el valor real de las posibles propiedades y decidir si invertir en ellas o no.

Para ese fin dispone de un conjunto de datos recopilados que contiene diferentes variables que describen los bienes y el precio de venta en cada caso. Con estos datos se busca elaborar un modelo estadístico que permita hacer predicciones del Precio de Venta.

2. Situación Problema.

En el competitivo mercado inmobiliario australiano, los precios de las propiedades pueden variar ampliamente en función de factores económicos, sociales y específicos del inmueble, como la ubicación, el tamaño, la antigüedad y el estado de mantenimiento. Para capitalizar estas variaciones de manera rentable, la empresa busca identificar propiedades cuyo valor actual esté por debajo de su valor de mercado real, lo que representa una oportunidad de compra estratégica.

Esta necesidad de predecir con precisión los precios surge de la intención de invertir únicamente en propiedades que, según un análisis, ofrecen el potencial de ser revendidas con una ganancia considerable. Con un conjunto de datos disponible de ventas y la aplicación de modelos estadísticos avanzados, la empresa pretende no solo evitar compras de alto riesgo, sino también optimizar su portafolio de inversiones para maximizar el retorno.

Por lo tanto, la predicción lo más precisa posible de precios es una herramienta esencial para detectar propiedades subvaloradas de forma sistemática, guiando la toma de decisiones hacia adquisiciones más informadas y financieramente ventajosas.

3. Objetivo General.

Desarrollar un modelo estadístico predictivo de regresión avanzada para estimar con mayor precisión el precio de venta de propiedades residenciales en el mercado australiano, con el fin de identificar oportunidades de inversión en las que los inmuebles se encuentren subvalorados con respecto a su valor real.

4. Objetivos Específicos.

- Analizar y limpiar el conjunto de datos para garantizar la calidad de los datos, abordando valores nulos, outliers, y variables redundantes, con el fin de construir una base sólida para el modelado predictivo.
- Identificar las características más influyentes en la variación de los precios de venta de las propiedades residenciales mediante técnicas de selección de variables y análisis de correlación, para optimizar la precisión del modelo predictivo.
- Desarrollar modelos de regresión aplicando técnicas de Lasso, Ridge, y ajustar sus hiperparámetros para reducir la varianza y mejorar la capacidad predictiva del modelo, evaluando su rendimiento mediante métricas adecuadas como el MSE (Error Cuadrático Medio) y el R^2 (Coeficiente de Determinación) en el conjunto de prueba.
- Plantear el mejor modelo de predicción de Precios de Ventas para bienes raíces de la empresa Surprise Housing en base a un análisis estadístico de un conjunto de datos existente.

5. Metodología.

5.1 Recopilación de datos.

El conjunto de datos para el estudio está compuesto por 1,460 registros y 81 columnas, la primera columna Id, es un serial numérico asignado a cada registro por lo que no se la toma en cuenta en el análisis.

Las 80 columnas restantes son las siguientes:

Tabla 1: Variables del Conjunto de datos Surprise Housing

No.	VARIABLE
1	MSSubClass: Identifies the type of dwelling involved in the sale.
2	MSZoning: Identifies the general zoning classification of the sale.
3	LotFrontage: Linear feet of street connected to property
4	LotArea: Lot size in square feet
5	Street: Type of road access to property
6	Alley: Type of alley access to property
7	LotShape: General shape of property
8	LandContour: Flatness of the property
9	Utilities: Type of utilities available
10	LotConfig: Lot configuration
11	LandSlope: Slope of property
12	Neighborhood: Physical locations within Ames city limits
13	Condition1: Proximity to various conditions
14	Condition2: Proximity to various conditions (if more than one is present)
15	BldgType: Type of dwelling
16	HouseStyle: Style of dwelling

17	OverallQual: Rates the overall material and finish of the house
18	OverallCond: Rates the overall condition of the house
19	YearBuilt: Original construction date
20	YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)
21	RoofStyle: Type of roof
22	RoofMatl: Roof material
23	Exterior1st: Exterior covering on house
24	Exterior2nd: Exterior covering on house (if more than one material)
25	MasVnrType: Masonry veneer type
26	MasVnrArea: Masonry veneer area in square feet
27	ExterQual: Evaluates the quality of the material on the exterior
28	ExterCond: Evaluates the present condition of the material on the exterior
29	Foundation: Type of foundation
30	BsmtQual: Evaluates the height of the basement
31	BsmtCond: Evaluates the general condition of the basement
32	BsmtExposure: Refers to walkout or garden level walls
33	BsmtFinType1: Rating of basement finished area
34	BsmtFinSF1: Type 1 finished square feet
35	BsmtFinType2: Rating of basement finished area (if multiple types)
36	BsmtFinSF2: Type 2 finished square feet
37	BsmtUnfSF: Unfinished square feet of basement area
38	TotalBsmtSF: Total square feet of basement area
39	Heating: Type of heating
40	HeatingQC: Heating quality and condition
41	CentralAir: Central air conditioning
42	Electrical: Electrical system
43	1stFlrSF: First Floor square feet
44	2ndFlrSF: Second floor square feet
45	LowQualFinSF: Low quality finished square feet (all floors)
46	GrLivArea: Above grade (ground) living area square feet
47	BsmtFullBath: Basement full bathrooms
48	BsmtHalfBath: Basement half bathrooms
49	FullBath: Full bathrooms above grade
50	HalfBath: Half baths above grade
51	Bedroom: Bedrooms above grade (does NOT include basement bedrooms)
52	Kitchen: Kitchens above grade
53	KitchenQual: Kitchen quality
54	TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
55	Functional: Home functionality (Assume typical unless deductions are warranted)
56	Fireplaces: Number of fireplaces
57	FireplaceQu: Fireplace quality
58	GarageType: Garage location
59	GarageYrBlt: Year garage was built
60	GarageFinish: Interior finish of the garage
61	GarageCars: Size of garage in car capacity
62	GarageArea: Size of garage in square feet

63	GarageQual: Garage quality
64	GarageCond: Garage condition
65	PavedDrive: Paved driveway
66	WoodDeckSF: Wood deck area in square feet
67	OpenPorchSF: Open porch area in square feet
68	EnclosedPorch: Enclosed porch area in square feet
69	3SsnPorch: Three season porch area in square feet
70	ScreenPorch: Screen porch area in square feet
71	PoolArea: Pool area in square feet
72	PoolQC: Pool quality
73	Fence: Fence quality
74	MiscFeature: Miscellaneous feature not covered in other categories
75	MiscVal: $Value of miscellaneous feature
76	MoSold: Month Sold (MM)
77	YrSold: Year Sold (YYYY)
78	SaleType: Type of sale
79	SaleCondition: Condition of sale
80	SalePrice Precio de venta

Fuente: Elaboración propia - Bienes Raíces - Mercado Australiano, 2024

Análisis de la calidad de los datos:

✓ Se han identificado gran cantidad de datos faltantes, en el análisis de calidad:

```
LotFrontage      259
Alley           1369
MasVnrType         8
MasVnrArea         8
BsmtQual          37
BsmtCond          37
BsmtExposure      38
BsmtFinType1      37
BsmtFinType2      38
Electrical         1
FireplaceQu      690
GarageType        81
GarageYrBlt       81
GarageFinish      81
GarageQual        81
GarageCond        81
PoolQC          1453
Fence           1179
MiscFeature     1406
```

Considerando que las columnas con mayor cantidad de datos faltantes no llegan a ser relevantes para el análisis, se ha decidido descartarlas para el análisis, como ser:

- Alley Tipo de callejón de acceso a la propiedad

- FireplaceQu Calidad de la chimenea
- PoolQC Calidad de la piscina
- Fence Calidad de la valla
- MiscFeature Caract. Miscelánea no cubierta en otras Categ.

Adicionalmente en relación a variables categóricas por exceso de nulos, se decide no tomar en cuenta las siguientes variables: GarageFinish', 'BsmtFinType2', 'BsmtFinType1', 'BsmtExposure', 'GarageQual', 'GarageCond', 'GarageType', 'BsmtCond', 'BsmtQual'.

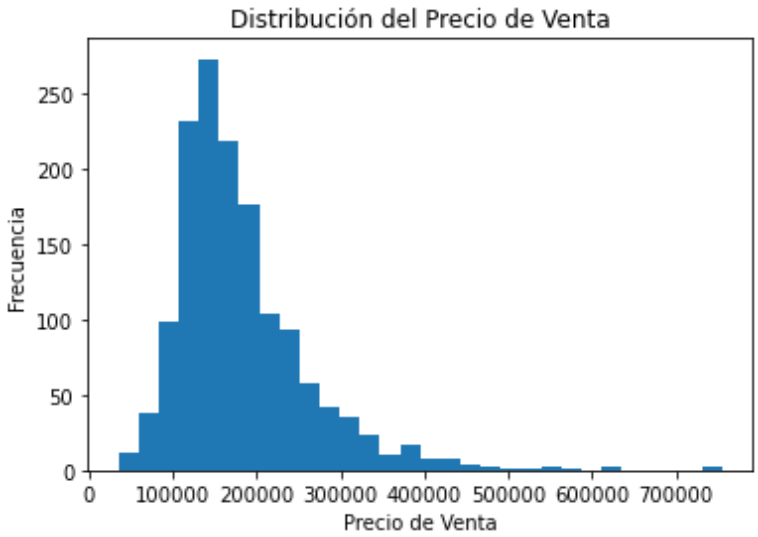
Para las otras columnas se han completado los datos faltantes considerando el contenido de la información recopilada:

- LotFrontage (Pies lineales de calle conectados a la propiedad). Convertir los valore nulos en 0.
- MasVnrType: Tipo de revestimiento de mampostería. Llenar los nulos con "Ninguno".
- MasVnrArea: Área de revestimiento de mampostería en pies cuadrados. Convertir los valore nulos en 0.
- Electrical : Sistema eléctrico. Reemplazaremos el nulo por Mixto.
- GarageYrBlt: Evalúa el garage: Los nulos significa que no tiene garaje. Llenar los nulos con "No".

5.2 Análisis Exploratorio de Datos (EDA).

➢ **Análisis de la variable objetivo: Salesprice**:

Gráfica 1: Distribución de la variable SalePrice

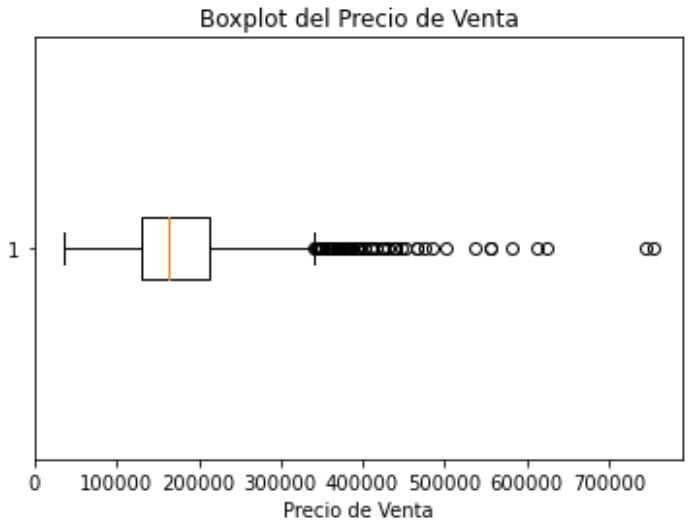


Fuente: Elaboración propia - Bienes Raíces - Mercado Australiano, 2024

Podemos ver que los precios están sesgados hacia la derecha y algunos valores extremos están por encima de aproximadamente 500,000. Tiene

una Media: 180,921.20 Desviación estándar: 79,442.50. Interpretación: SalePrice, que representa el precio de venta de las casas, tiene una alta dispersión. Esto indica una gran variabilidad en los precios de las viviendas, probablemente debido a diferencias en el tamaño, la calidad, y las características de las propiedades.

Grafica 2: Grafico de caja de la variable SalePrice



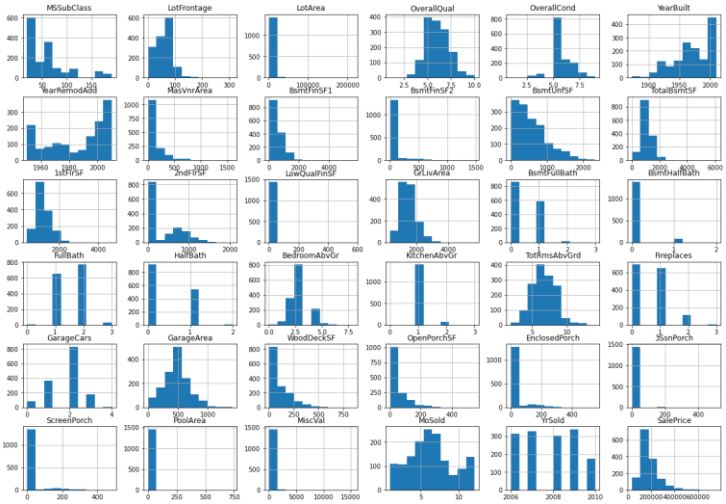
Fuente: Elaboración propia - Bienes Raíces - Mercado Australiano, 2024

El diagrama de cajas muestra claramente lo que inicialmente transmiten los histogramas, que existe gran dispersión en la variable. Existen precios muy altos en comparación con precios muy bajos.

- **Análisis de variables numéricas:**

A continuación, se revisan los histogramas de distribución de cada variable:

Gráfico 3: Histogramas de distribución de variables



Fuente: Elaboración propia - Bienes Raíces - Mercado Australiano, 2024

Sesgo Positivo: Variables como LotArea, TotalBsmtSF, y GarageArea parecen tener un sesgo positivo, donde la mayor parte de los datos se agrupan hacia la izquierda con algunos valores extremos a la derecha. Esto sugiere que no se distribuyen normalmente.

Sesgo Negativo: Variables como Fireplaces y MiscVal muestran una concentración de valores en la parte inferior, sugiriendo un sesgo negativo.

Distribuciones Bimodales o Multimodales: Algunas variables, como OverallQual y TotRmsAbvGrd, pueden presentar características bimodales, indicando que podría haber subgrupos en los datos.

Concentraciones en Valores Específicos: Variables como MoSold y SalePrice pueden mostrar picos específicos, lo que sugiere que ciertos rangos de precios o meses de venta pueden ser más frecuentes.

Normalidad en Distribuciones: Algunas variables pueden aproximarse a una normalidad, especialmente aquellas que parecen tener una forma más simétrica, pero la mayoría de las variables en los histogramas presentados parecen desviarse de una distribución normal.

✓ Análisis de Correlación:

SalePrice	1.00
OverallQual	0.79
GrLivArea	0.71
GarageCars	0.64
GarageArea	0.62
TotalBsmtSF	0.61
1stFlrSF	0.61
FullBath	0.56
TotRmsAbvGrd	0.53
YearBuilt	0.52
YearRemodAdd	0.51
MasVnrArea	0.47
Fireplaces	0.47
BsmtFinSF1	0.39
WoodDeckSF	0.32
2ndFlrSF	0.32
OpenPorchSF	0.32
HalfBath	0.28
LotArea	0.26
BsmtFullBath	0.23
BsmtUnfSF	0.21
LotFrontage	0.21
BedroomAbvGr	0.17
ScreenPorch	0.11
PoolArea	0.09
MoSold	0.05

Las variables con las correlaciones más altas con SalePrice son: OverallQual (0.79): La calidad general de la casa es el factor más fuerte que influye en el precio de venta. GrLivArea (0.71): El área habitable sobre el suelo también tiene un impacto significativo en el precio. GarageCars

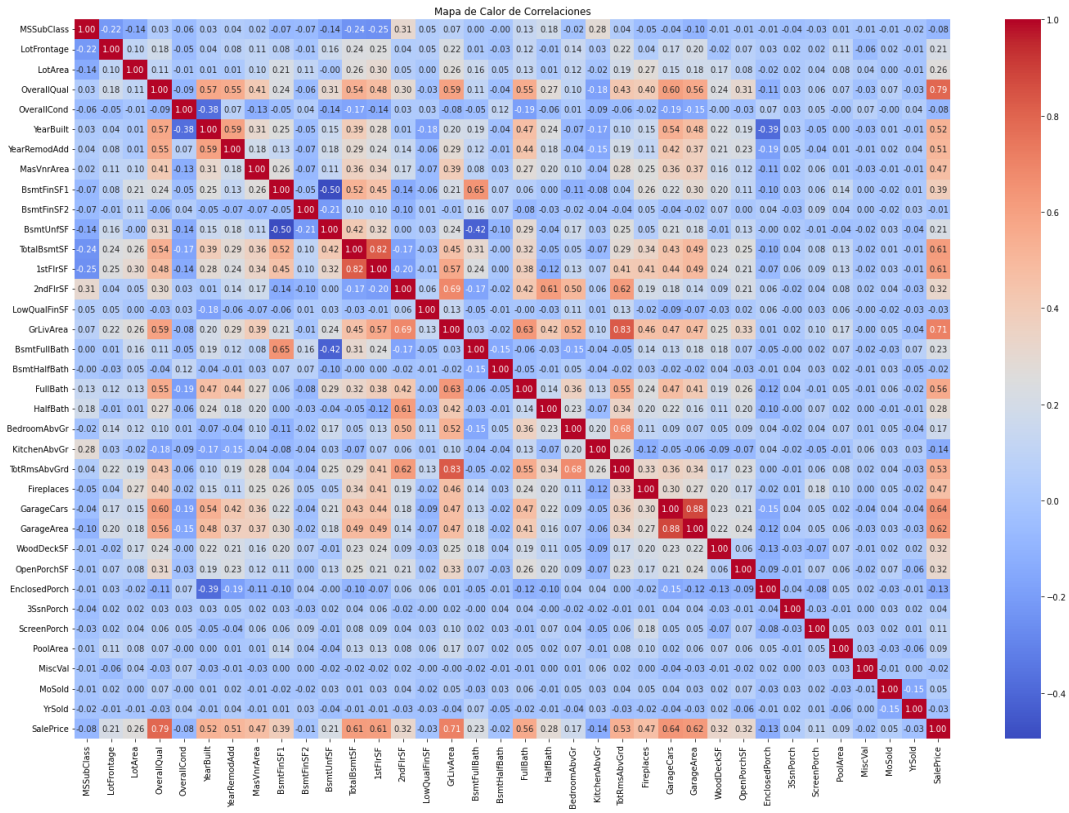
(0.64): La cantidad de espacios en el garaje es un fuerte predictor del precio. GarageArea (0.62): El área del garaje también muestra una correlación considerable. TotalBsmtSF (0.61) y 1stFlrSF (0.61): Ambas dimensiones del área del sótano y del primer piso son importantes.

Correlación Moderada: Variables como FullBath (0.56) y TotRmsAbvGrd (0.53) (total de habitaciones sobre el nivel del suelo) también tienen una correlación moderada, lo que indica que más baños y habitaciones tienden a asociarse con precios más altos.

Correlaciones Menores: Variables como MoSold (0.05) (mes de venta) y PoolArea (0.09) tienen correlaciones bastante bajas, lo que sugiere que estos factores son menos relevantes para el precio de venta.

Variables Irrelevantes: Variables como BsmtFinSF2 (-0.01) y MiscVal (-0.02) presentan correlaciones cercanas a cero, lo que sugiere que no tienen un impacto significativo en el precio de venta.

Grafica 4: Mapa de calor de Correlaciones de variables



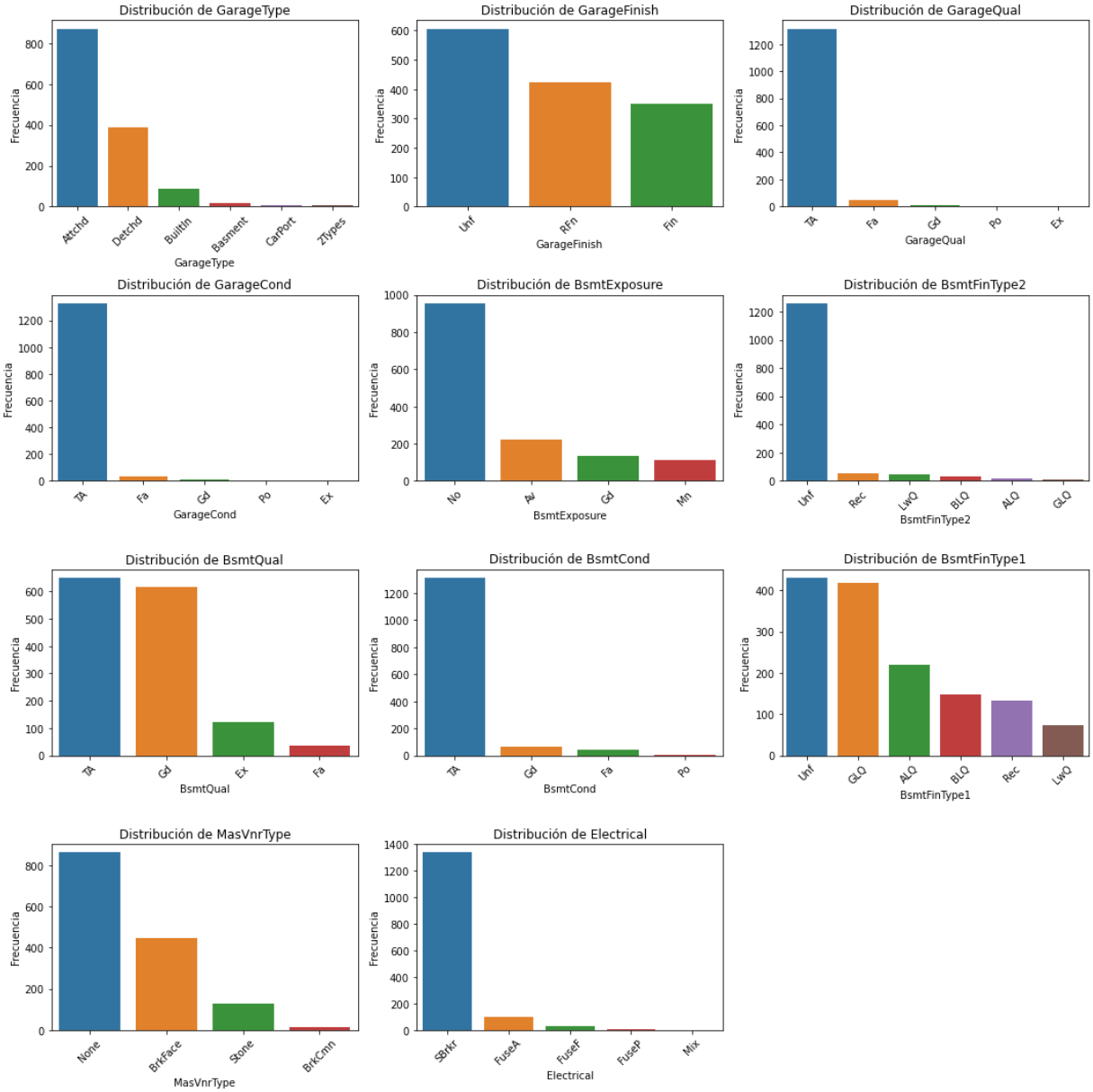
Fuente: Elaboración propia - Bienes Raíces - Mercado Australiano, 2024

✓ Análisis de valores atípicos, se realiza reemplazo de valores atípicos, calculando para ese fin los cuartiles y rango intercuartílico de cada variable, esto consecuentemente mejora el valor de correlación de las variables.

➢ **Análisis de variables categóricas:**

Gráficos de barras variables categorías:

Grafica 5: Histogramas de distribución de variables categóricas



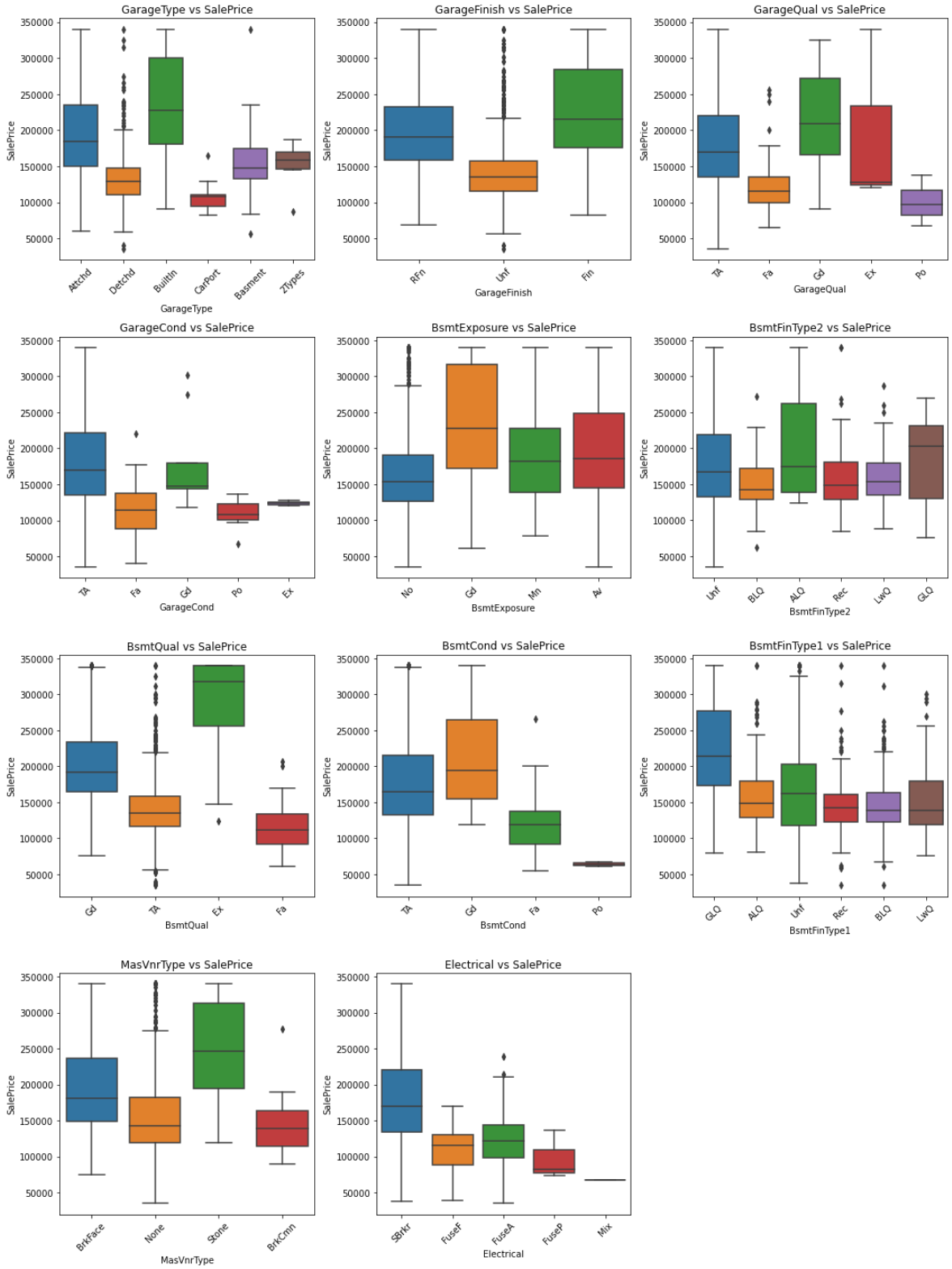
Fuente: Elaboración propia - Bienes Raíces - Mercado Australiano, 2024

Interpretación

1. **GarageType**: La mayoría de los garajes son de tipo "Attached", con un número considerable en "Detached", mientras que otros tipos como "BuiltIn" y "CarPort" son bastante menos frecuentes.
2. **GarageFinish**: La distribución muestra que el tipo de acabado de garaje más común es "Unf", seguido por "RFn" y "Fin". Esto sugiere que muchos garajes no son terminados.

3. **GarageQual**: La calidad del garaje está predominantemente en la categoría "TA" (Typical/Average), con muy pocos datos en las categorías de calidad superior o inferior.
4. **GarageCond**: La condición de los garajes es mayoritariamente "TA", lo que implica que la mayoría se encuentra en condiciones normales, mientras que las condiciones extremas son muy raras.
5. **BsmtExposure**: La exposición del sótano muestra que la mayoría de los sótanos tienen poco o ninguna exposición ("No" y "Av"), con muy pocos casos que tienen una exposición más significativa.
6. **BsmtQual**: Similar al patrón anterior, la calidad del sótano es mayoritariamente "TA", indicando una calidad promedio en la mayoría de los sótanos.
7. **BsmtCond**: La condición del sótano también muestra una predominancia en "TA", con pocas observaciones en otras categorías.
8. **BsmtFinType1**: Esta variable muestra una gran cantidad de "Unf" (No terminado), con algunas proporciones más pequeñas en otras categorías, lo que sugiere que muchos sótanos no están terminados.
9. **BsmtFinType2**: Al igual que con la primera fin de sótano, la mayoría son "Unf", pero hay una diversidad mayor en otras categorías.
10. **MasVnrType**: La mayoría de las casas tienen "None" como tipo de revestimiento, seguido de "BrkFace" y "Stone", lo que sugiere que la gran mayoría no tienen revestimiento de mampostería.
11. **Electrical**: La gran mayoría de las casas tienen un sistema eléctrico que se clasifica como "SBrkr", lo que implica que existe un estándar moderno predominante en la instalación eléctrica.

Gráfica 6: Diagrama de cajas de variables categóricas



Fuente: Elaboración propia - Bienes Raíces - Mercado Australiano, 2024

En general, estos gráficos sugieren que existen ciertas tendencias predominantes en los tipos y calidades de las características residenciales

representadas, con una mayoría de casas mostrando condiciones y acabados promedio.

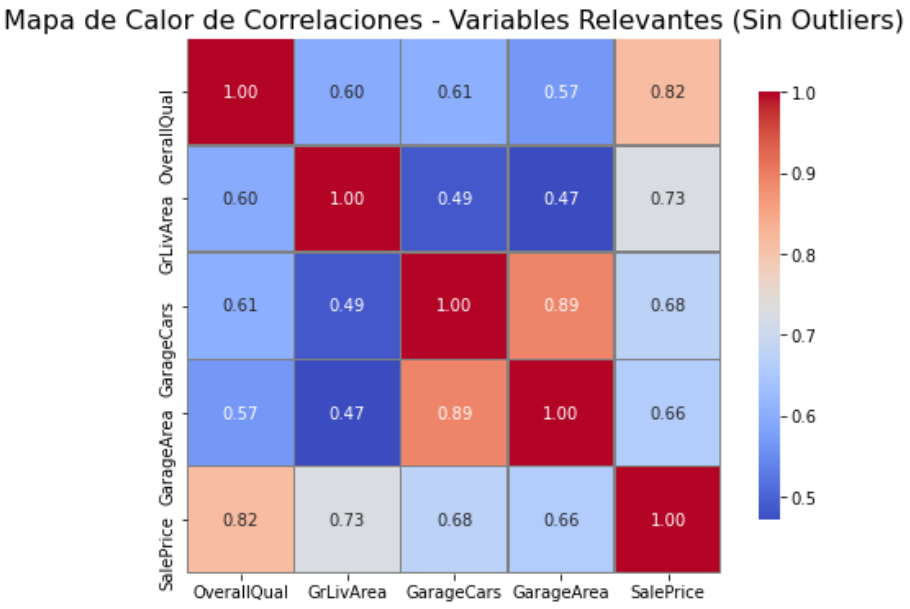
Se procede al tratamiento de outliers: se ha realizado un agrupamiento de categorías raras y tratar outliers en las variables de esta manera se ayuda a mejorar la calidad del modelo y su interpretabilidad. Las categorías con pocas observaciones pueden causar inestabilidad y ruido en el modelo, por lo que se agrupan bajo etiquetas comunes como “Otro” o “Raro”, reduciendo así la complejidad del análisis.

Este enfoque simplifica el conjunto de datos, ayuda a evitar el sobreajuste y mejora la eficiencia del modelo al reducir la dimensionalidad. Además, la agrupación de categorías infrecuentes permite mantener la integridad de los datos sin eliminar información valiosa, resultando en un modelo más robusto y fácil de interpretar.

5.3 Selección de variables.

- ✓ Variables numéricas seleccionadas para el análisis: concluido el análisis exploratorio con relación a las variables numéricas y considerando los índices de correlación, se seleccionan las siguientes variables a considerar para el estudio:'OverallQual', 'GrLivArea', 'GarageCars', 'GarageArea':

Gráfica 7: Mapa de calor de variables relevantes numéricas



Fuente: Elaboración propia - Bienes Raíces - Mercado Australiano, 2024

- ✓ Luego del análisis de valores nulos y outliers, las variables categóricas seleccionadas finalmente para el análisis son:

```
Categorías en 'MSZoning':
```

```
Categorías en 'Street':
Categorías en 'LotShape':
Categorías en 'LandContour':
Categorías en 'Utilities':
Categorías en 'LotConfig':
Categorías en 'LandSlope':
Categorías en 'Neighborhood':
Categorías en 'Condition1':
Categorías en 'Condition2':
Categorías en 'BldgType':
Categorías en 'HouseStyle':
Categorías en 'RoofStyle':
Categorías en 'RoofMatl':
Categorías en 'Exterior1st':
Categorías en 'Exterior2nd':
Categorías en 'MasVnrType':
Categorías en 'ExterQual':
Categorías en 'ExterCond':
Categorías en 'Foundation':
Categorías en 'Heating':
Categorías en 'HeatingQC':
Categorías en 'CentralAir':
Categorías en 'Electrical':
Categorías en 'KitchenQual':
Categorías en 'Functional':
Categorías en 'PavedDrive':
Categorías en 'SaleType':
Categorías en 'SaleCondition':
```

6. Desarrollo de Modelos Estadísticos.

6.1 Descripción de los Modelos Propuestos.

- **Modelo de Regresión Lineal con Variables numéricas.**

  En el desarrollo del primer modelo estadístico de Regresión Lineal para la predicción de precios de propiedades (SalesPrice), se han seleccionado como variables predictoras cuatro atributos clave: OverallQual (calidad general de la construcción), GrLivArea (área habitable por encima del suelo), GarageCars (capacidad del garaje en número de autos), y GarageArea (área total del garaje en pies cuadrados). Estos atributos se identificaron como los factores con mayor influencia en el precio de venta de una propiedad, y el modelo ha sido diseñado para capturar su relación directa y lineal con el precio.

- **Modelo de Regresión considerando las variables numéricas y categóricas fundamentales.**

  Se realiza un tratamiento de las variables categóricas seleccionadas para convertirlas en variables dummy.

  Se procede a combinar las variables numéricas con mayor índice de correlación con nuestra variable objetivo, junto a las variables categóricas seleccionadas, el set de datos será utilizado para elaborar un analisis de regresión mediante la función de LinearRegression de la librería sklearn. La regresión lineal es un modelo predictivo que intenta establecer una relación lineal entre las variables independientes (características de la casa) y la variable dependiente (precio de venta).

- **Aplicando Ridge al Modelo de Regresión.**

  Aplicar Ridge en el análisis de regresión implica agregar un término de regularización a la función de costo del modelo, que penaliza los valores grandes en los coeficientes de las variables predictoras. Esto se hace con el objetivo de reducir el riesgo de sobreajuste (*overfitting*), especialmente cuando se tienen múltiples variables predictoras o variables altamente correlacionadas (Hastie et al., 2009).

  En Ridge, el término de regularización es proporcional a la suma de los cuadrados de los coeficientes, y está controlado por el parámetro alpha: un valor más alto de alpha da mayor penalización y reduce el tamaño de los coeficientes, mientras que un valor bajo reduce el impacto de la regularización. Para el análisis se ha tomado un valor de Alpha de 1.

- **Aplicando Lasso al Modelo de Regresión.**

- El modelo de regresión Lasso (*Least Absolute Shrinkage and Selection Operator*) es una técnica de regularización aplicada en modelos de regresión lineal, cuyo principal objetivo es mejorar la precisión de predicción y simplificar el modelo al reducir la magnitud de los coeficientes de las variables predictoras. Esto se logra penalizando los coeficientes en función de su magnitud, eliminando aquellos menos relevantes (reduciéndolos a cero) y reteniendo solo las variables más significativas (James et al., 2013).

- La regularización Lasso es particularmente útil en conjuntos de datos con muchas variables, como podría interpretarse en nuestro caso de estudio, ya que realiza automáticamente una selección de características, lo cual puede resultar en un modelo más interpretable y menos susceptible a sobreajuste. Para el análisis se ha tomado un valor de Alpha de 0.1

- **Aplicando ElasticNet al Modelo de Regresión.**

ElasticNet es un modelo de regresión lineal que combina las penalizaciones de Lasso y Ridge mediante la regularización, lo cual permite obtener un equilibrio entre los beneficios de ambos métodos. ElasticNet es especialmente útil cuando hay muchas variables y algunas de ellas están correlacionadas entre sí, ya que:

Lasso (L1) tiene la ventaja de realizar selección de variables, reduciendo algunos coeficientes a cero, lo que ayuda a simplificar el modelo.

Ridge (L2) minimiza el impacto de las multicolinealidades al reducir la magnitud de los coeficientes de las variables correlacionadas (James et al., 2013).

ElasticNet utiliza dos hiperparámetros clave:

alpha: controla la magnitud de la penalización general. A mayor valor de alpha, mayor es la penalización, lo que reduce el sobreajuste.

l1_ratio: define el balance entre Lasso y Ridge. Un valor de 0.5 significa que ambas penalizaciones se ponderan de manera igual.

6.2 Entrenamiento y Validación.

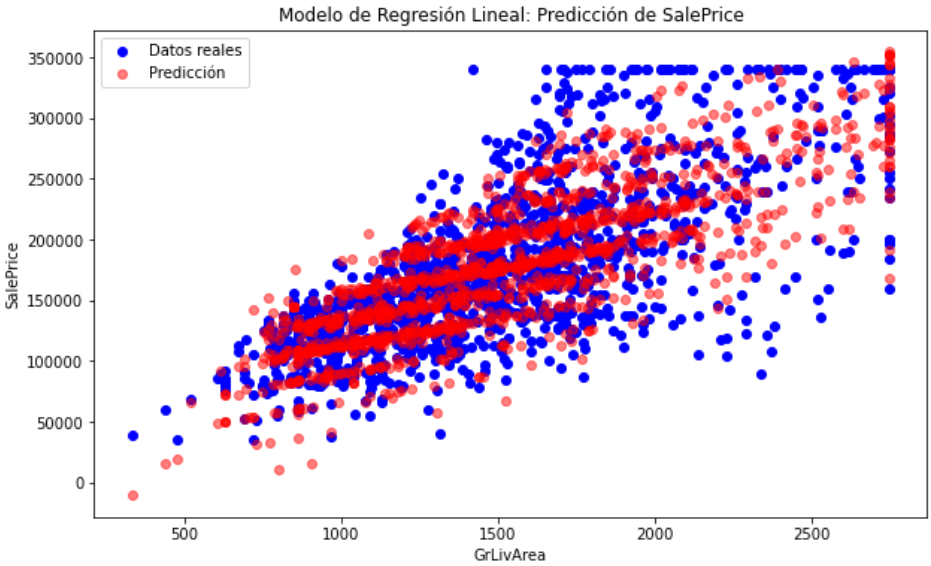
- **Modelo de Regresión Lineal con Variables numéricas.**

  Entrenamiento y prueba: para el análisis dividimos los datos de tal forma que el 20% de los datos se utilizará para el conjunto de prueba, dejando el 80% para el entrenamiento y random_state=42 asegura que la división sea reproducible.

  Métricas del modelo:

  - RMSE (Error Cuadrático Medio de Raíz): 28,933.35 sugiere que, en promedio, las predicciones del modelo de precio de venta (SalesPrice) se desvían del valor real en aproximadamente 28,933 unidades de la moneda utilizada.
  - R^2 (Coeficiente de Determinación): 0.8288, el modelo explica aproximadamente el 82.88% de la variabilidad total en los precios de venta de las propiedades. Esto indica que las variables seleccionadas (OverallQual, GrLivArea, GarageCars y GarageArea) son efectivas para predecir el precio de venta. Sin embargo, el 17.12% restante de la variabilidad en el precio no es explicado por el modelo.

Gráfica 8: Modelo de regresión lineal Predicción de SalePrice



Fuente: Elaboración propia - Bienes Raíces - Mercado Australiano, 2024

- **Modelo de Regresión considerando las variables numéricas y categóricas fundamentales.**

  Entrenamiento y prueba: para el análisis mantenemos los mismos parámetros que el caso de las variables numéricas solas, dividimos los datos de tal forma que el 20% de los datos se utilizará para el conjunto de prueba, dejando el 80% para el entrenamiento y random_state=42 asegura que la división sea reproducible.

  Métricas del modelo:

  - RMSE (Error Cuadrático Medio de Raíz): con un RMSE de 23,294.81, sugiere que, en promedio, las predicciones del modelo de precio de venta (SalesPrice) se desvían del valor real en aproximadamente 23,294 unidades de la moneda utilizada.

  - R^2 (Coeficiente de Determinación): el R^2 de 0.8890 indica que el modelo explica ahora el 88.90% de la variabilidad en el precio de venta. Esto indica que las variables seleccionadas numéricas más las categorías son efectivas para predecir el precio de venta. Sin embargo, el 11.1% restante de la variabilidad en el precio no es explicado por el modelo.

6.3 Análisis de multicolinealidad - Método del Factor Inflador de Varianza (VIF).

El Análisis de Multicolinealidad mediante el Factor Inflador de Varianza (VIF) es una técnica que permite identificar y cuantificar la presencia de

multicolinealidad entre las variables independientes en un modelo de regresión. La multicolinealidad ocurre cuando dos o más variables predictoras están altamente correlacionadas, lo que puede llevar a problemas en la interpretación de los coeficientes del modelo y a un sobreajuste, afectando la precisión de las predicciones.

¿Qué es el VIF?
El Factor Inflador de Varianza (VIF) mide cuánto aumenta la varianza (incertidumbre) de los coeficientes de una variable en función de su correlación con otras variables en el modelo. Se calcula para cada variable independiente y permite evaluar el grado de multicolinealidad.

Para el análisis realizado se ha identificado:

Variables con Alto VIF
OverallQual: 89.23

Este valor es extremadamente alto, lo que indica que OverallQual tiene una alta correlación con una o más de las otras variables en el modelo. Esto puede causar inestabilidad en los coeficientes de la regresión y sugerir que se podría considerar eliminar o transformar esta variable para mejorar la interpretación del modelo.

GrLivArea: 48.00

También tiene un VIF alto, lo que sugiere que está correlacionada con otras variables. GrLivArea (área habitable) puede ser una variable importante, pero deberías revisar qué otras variables están causando esta multicolinealidad.

GarageCars: 46.66

Similar a GrLivArea, sugiere que GarageCars tiene relaciones significativas con otras variables que podrían estar afectando la calidad del modelo.

GarageArea: 41.24

Al igual que las anteriores, este alto VIF indica multicolinealidad, lo que puede hacer que el modelo sea menos confiable.

MSZoning_FV: 16.18

Aunque el VIF es alto, está por debajo de 20. Esto indica que, aunque puede haber correlación, no es tan severa como en los casos anteriores. Sin embargo, es una señal para revisar su relación con otras variables.

Variables con Bajo VIF

SaleCondition_Alloca: 1.49 SaleCondition_Family: 1.34 SaleCondition_Otros: 1.42 SaleCondition_Partial: 50.31

Estas variables tienen VIF cercanos a 1, lo que indica que no tienen multicolinealidad y que son independientes de las otras variables en el modelo. Esto es deseable.

Identificar y Tratar la Multicolinealidad: Las variables con VIF mayores a 10 (como OverallQual, GrLivArea, GarageCars, GarageArea) sugieren que hay multicolinealidad significativa en el modelo. Sería prudente considerar eliminar o combinar algunas de estas variables.

Revisar Relaciones entre Variables: Se debe investigar las correlaciones entre las variables con alto VIF para determinar cuáles pueden estar causando la multicolinealidad. Puedes hacerlo utilizando un mapa de calor (heatmap) para visualizar las correlaciones.

Modelos de Regularización: Si decides mantener estas variables, considera usar métodos de regularización (como Ridge Regression o Lasso Regression) que pueden ayudar a manejar la multicolinealidad al penalizar los coeficientes de las variables correlacionadas.

Ajustar el Modelo: Basado en la interpretación de los VIF, puedes ajustar el modelo para mejorar su estabilidad y predicción.

- **Aplicando Ridge al Modelo de Regresión con corrección de Multicolinealidad.**

Conclusiones de multicolinealidad de modelo Ridge:

Variable const: VIF de 2606.50
Este valor es extremadamente alto y sugiere problemas serios de multicolinealidad. Esto se debe a que el término constante puede no tener una interpretación directa, pero indica que al menos una variable tiene una alta correlación con otras variables. Otras Variables y sus VIFs

OverallQual (4.45), GrLivArea (4.56), GarageCars (7.10), GarageArea (6.70): Estas variables tienen VIFs superiores a 4, lo que indica que hay cierta multicolinealidad presente, pero no es necesariamente crítica. Los valores de VIF entre 4 y 10 indican que las variables están correlacionadas, lo que puede ser problemático, pero no al nivel más severo. SaleCondition_Partial (46.04): Este valor es muy alto, lo que indica que esta variable tiene una fuerte colinealidad con otras variables del modelo. Se recomienda evaluar esta variable en particular, ya que podría estar influyendo en la estabilidad del modelo.

Métricas finales del modelo:

Ridge RMSE (Root Mean Squared Error): 32,445.60

El error promedio en las predicciones del modelo Ridge, indica que, en promedio, las predicciones de precios de venta (SalesPrice) se desvían de los valores reales en aproximadamente 32,445.60 unidades monetarias.

Ridge R^2: 0.7667

El valor de R^2 = 0.7667 indica que el modelo Ridge explica el 76.67% de la variabilidad en los precios de venta de las propiedades en el conjunto de prueba.

- **Aplicando Lasso al Modelo de Regresión con corrección de Multicolinealidad.**

Para verificar la multicolinealidad en el modelo Lasso, puedes calcular el VIF (Variance Inflation Factor) de las variables en el conjunto de datos, similar a como se hace en otros modelos de regresión lineal. Sin embargo, dado que Lasso (a diferencia de Ridge) tiende a reducir los coeficientes de variables menos relevantes a cero, el efecto de la multicolinealidad se mitiga en cierta medida de manera automática. Aun así, calcular el VIF antes de ajustar el modelo puede ayudarte a entender las relaciones entre las variables.
Métricas del modelo:

Lasso RMSE (Root Mean Squared Error) = 23,289.57

El RMSE obtenido con el modelo Lasso es 23,289.57, lo que indica el error promedio de las predicciones en términos de la desviación con respecto a los valores reales. Un RMSE bajo es indicativo de un modelo preciso y representa la eficacia del modelo en predecir valores cercanos a los reales.

Lasso R^2 = 0.8891
El R^2 obtenido es 0.8891, lo que indica que el modelo Lasso explica aproximadamente el 88.91% de la variabilidad en los precios de venta (SalesPrice). Esto implica que las variables seleccionadas tras la regularización capturan muy bien la relación con el precio, manteniendo una capacidad predictiva alta.

- **Aplicando ElasticNet al Modelo de Regresión con corrección de Multicolinealidad.**

Métricas del modelo:

ElasticNet RMSE: 22,944.54

Este valor representa el Error Cuadrático Medio de Raíz (RMSE), indicando que, en promedio, las predicciones del modelo se desvían de los valores reales en aproximadamente 22,944 unidades de la moneda.

ElasticNet R^2: 0.8923

El R^2 de 0.8923 implica que el modelo explica el 89.23% de la variabilidad en el precio de venta de las propiedades.

6.4 Selección del Modelo Final.

- **Modelo de Regresión Lineal con Variables numéricas.**

  El modelo tiene un buen ajuste con un R2 superior al 82%, indicando que casi el 80% de la variabilidad en el precio de las casas puede ser explicada por las variables OverallQual, GrLivArea, GarageCars, y GarageArea. La prueba F indica que al menos una de estas variables tiene un efecto significativo en el precio de venta. Este modelo parece prometedor, pero siempre es recomendable realizar diagnósticos adicionales (como verificar supuestos de residuos) para asegurar su validez.

- **Modelo de Regresión considerando las variables numéricas y categóricas fundamentales.**

  Incluir las variables categóricas en el modelo de regresión lineal ha mejorado los resultados de precisión, con un RMSE de 23,294.81, el error promedio de las predicciones se ha reducido en comparación con el modelo anterior (que tenía un RMSE de 28,933.35). Esto significa que el modelo con variables categóricas está haciendo predicciones de SalesPrice (precio de venta) con menos desviación respecto a los valores reales, aumentando así su precisión.

  Adicionalmente, el Coeficiente de Determinación R^2 de 0.8890 indica que el modelo explica ahora el 88.90% de la variabilidad en el precio de venta, una mejora notable respecto al 82.88% obtenido en el modelo sin variables categóricas. Esto sugiere que las variables categóricas agregadas tienen una relación significativa con el precio de venta, permitiendo capturar mejor las diferencias en las características de las propiedades.

  En conclusión, la inclusión de variables categóricas ha hecho que el modelo sea más robusto, aumentando su capacidad para hacer predicciones más precisas y explicativas en relación con el precio de venta de las propiedades.

- **Aplicando Ridge al Modelo de Regresión.**

  El valor de Ridge RMSE (Root Mean Squared Error) de 32,445.60 en comparación con el modelo sin regularización (RMSE de 23,294.81), el modelo Ridge no ha logrado una reducción en el error de predicción, sin embargo se debe considerar la corrección de multicolinealidad.

El valor de R^2 = 0.7667 indica que el modelo Ridge explica el 76.67% de la variabilidad, esto no representa una mejora en comparación con el modelo anterior (R^2 de 0.8890), sugiriendo que la regularización Ridge no ha optimizado el ajuste.

- **Aplicando Lasso al Modelo de Regresión.**

RMSE (Root Mean Square Error): 23,289.57

Este valor representa el error cuadrático medio de las predicciones del modelo sobre el conjunto de prueba. Cuanto menor sea este valor, más preciso será el modelo. En este caso, un RMSE de aproximadamente 23,289 indica que el modelo Lasso tiene un error moderado, lo cual podría mejorarse reduciendo variables redundantes o afinando el modelo. R^2 (Coeficiente de Determinación): 0.889

El valor de $R2$ indica que el modelo explica el 88.9% de la variabilidad en el precio de venta (SalePrice), lo cual es bastante bueno. Interpretación del impacto de Lasso:

El modelo Lasso, al penalizar los coeficientes de las variables menos relevantes, ayuda a reducir la complejidad y la multicolinealidad en el modelo. Esto se puede observar si algunos coeficientes en lasso_coefficients se han reducido a cero, lo que indica que Lasso eliminó algunas variables. Aunque el modelo Lasso puede perder algo de capacidad predictiva en comparación con Ridge (ligera disminución en R2), este efecto es compensado por una mayor simplicidad y menor riesgo de multicolinealidad. En conclusión, el modelo ha mejorado.

- **Aplicando ElasticNet al Modelo de Regresión.**

El valor de ElasticNet RMSE: 22,944.54 (Error Cuadrático Medio de Raíz) muestra que el modelo ElasticNet mejora ligeramente en precisión respecto a modelos previos, al reducir el RMSE.

El valor de ElasticNet R^2: 0.8923 refleja una mejora en la capacidad explicativa del modelo, posiblemente debida a la combinación de Lasso y Ridge, que optimiza la influencia de las variables relevantes y reduce el ruido de variables menos útiles. Comparativamente seria el mejor modelo para predicción del estudio.

Aplicar ElasticNet en el análisis permite obtener un modelo que es más preciso y explicativo, con un error reducido y un mayor poder predictivo, gracias a la combinación de técnicas de regularización que mejoran su robustez. ElasticNet, al combinar los beneficios de Lasso y Ridge, puede adaptarse mejor a datos complejos, ofreciendo una solución balanceada entre la selección de variables y la reducción de la multicolinealidad.

7. Resultados y Selección del modelo.

A continuación, un cuadro resumen de las métricas obtenidas de los modelos planteados en el análisis:

Tabla 2: Cuadro Comparativo de Métricas

CUADRO COMPARATIVO DE METRICAS DE AJUSTE DE MODELOS DE REGRESION		
MODELO DE REGRESION	RMSE (Error Cuadrático Medio de Raíz)	R^2 (Coeficiente de Determinación)
Regresión Lineal con Variables numéricas	28,933.35	0.8288
Regresión con variables numéricas y categóricas	23,294.81	0.8890
Ridge	32,445.60	0.7667
Lasso	23,289.57	0.8891
ElasticNet	22,944.54	0.8923

Fuente: Elaboración propia - Bienes Raíces - Mercado Australiano, 2024

El modelo estadístico de predicción elegido para el análisis es LASSO.

Eficacia del Modelo: El modelo Lasso ha proporcionado un buen ajuste con un RMSE relativamente bajo y un R^2 alto, lo que indica que las variables seleccionadas son efectivas para predecir la variable objetivo.

Impacto de las Variables: Puedes identificar las variables más influyentes en tu modelo (positivas y negativas) para entender mejor qué factores contribuyen al valor de la variable objetivo. Esto puede ser útil para la toma de decisiones o para realizar análisis más profundos sobre el conjunto de datos.

Multicolinealidad: Dado que estás utilizando Lasso, que es útil para manejar la multicolinealidad al penalizar las magnitudes de los coeficientes, esto puede ayudar a simplificar el modelo al eliminar variables menos significativas.

En resumen, los resultados sugieren que el modelo Lasso es efectivo en la predicción de la variable objetivo, y puedes usar los coeficientes para comprender la relación entre las variables predictivas y la variable objetivo.

7.1 Rendimiento del Modelo Final.

Lasso RMSE: 23256.79

Interpretación: El RMSE (Root Mean Squared Error) indica la cantidad promedio de error en las predicciones del modelo Lasso. Un RMSE de 23256.79 significa que, en promedio, las predicciones del modelo se desvían en esta cantidad de la variable objetivo real (en este caso, probablemente el precio de una casa, si estamos hablando de un conjunto de datos de vivienda). Este valor es relativamente bajo, lo que sugiere que el modelo realiza predicciones bastante precisas.

Lasso R^2: 0.8894

Interpretación: El R^2 (coeficiente de determinación) mide la proporción de la varianza en la variable dependiente que es predecible a partir de las variables independientes. Un R^2 de aproximadamente 0.8894 indica que el modelo Lasso explica alrededor del 88.94% de la variación en los datos. Esto es un buen resultado, sugiriendo que el modelo tiene un buen ajuste y que la mayoría de las variaciones en los precios (o la variable objetivo) son capturadas por las variables incluidas en el modelo.

Coeficientes del Modelo Lasso

Los coeficientes son cruciales porque indican la influencia de cada variable en la predicción de la variable objetivo. Vamos a desglosar esto:

Variables Positivas (Ejemplo: MSZoning_FV, SaleCondition_Alloca, SaleType_Otros, Neighborhood_StoneBr, Neighborhood_Veenker):

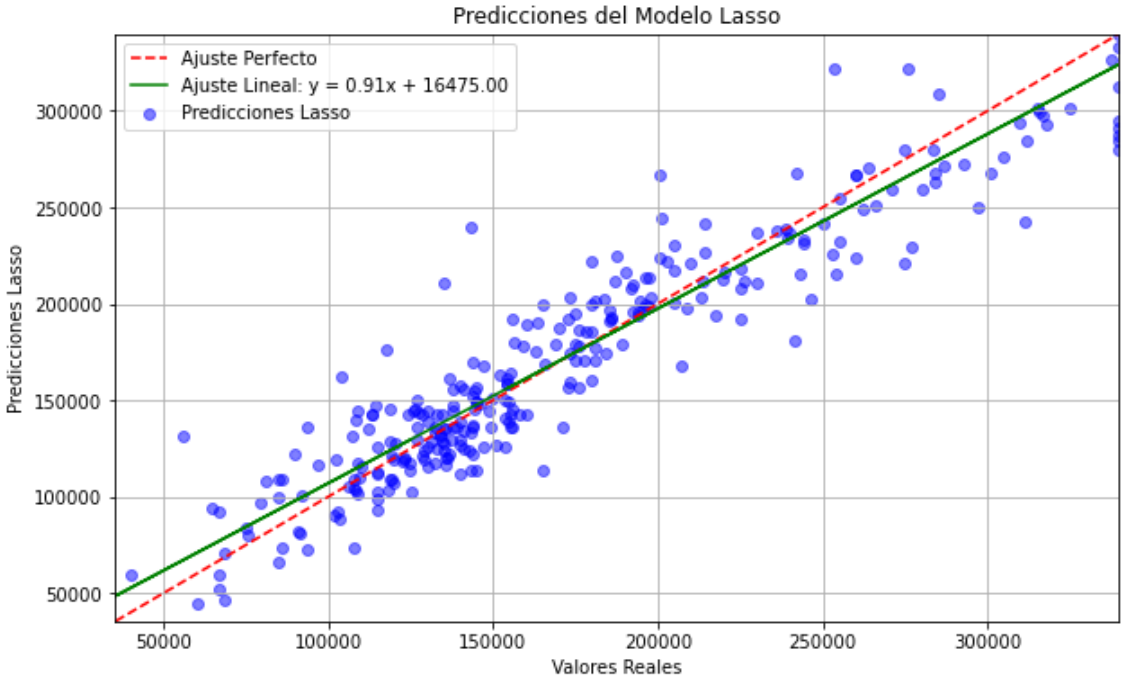
Interpretación: Estas variables tienen coeficientes positivos, lo que significa que un aumento en estas variables está asociado con un aumento en la variable objetivo (por ejemplo, el precio de la casa). Por ejemplo, un coeficiente de 38962.55 para MSZoning_FV indica que, manteniendo todas las demás variables constantes, un aumento en esta variable se asocia con un aumento de aproximadamente 38962.55 unidades en la variable objetivo.

Variables Negativas (Ejemplo: RoofStyle_Hip, RoofStyle_Gable, RoofStyle_Gambrel, Functional_Otros, RoofMatl_Otros):

Interpretación: Estas variables tienen coeficientes negativos, lo que indica que un aumento en estas variables se asocia con una disminución en la variable objetivo. Por ejemplo, un coeficiente de -74554.60 para RoofMatl_Otros indica que, manteniendo todas las demás variables constantes, un aumento en esta variable se asocia con una disminución de aproximadamente 74554.60 unidades en la variable objetivo. Esto sugiere que ciertas condiciones de techo pueden estar asociadas con precios más bajos.

7.2 Visualización de Resultados.

Grafica 9: Predicciones del modelo LASSO



Fuente: Elaboración propia - Bienes Raíces - Mercado Australiano, 2024

Modelo Lasso:

y = 0 + 38962.55 * MSZoning_FV + 38699.30 * SaleCondition_Alloca + 37688.58 * SaleType_Otros + 34020.19 * Neighborhood_StoneBr + 33686.99 * Neighborhood_Veenker + -47018.85 * RoofStyle_Hip + -47788.84 * RoofStyle_Gable + -47913.76 * RoofStyle_Gambrel + -49267.91 * Functional_Otros + -74554.60 * RoofMatl_Otros

El modelo Lasso ha demostrado ser efectivo en la predicción de precios, alcanzando un RMSE de aproximadamente 23,257 y un R^2 de 0.889, lo que indica que alrededor del 88.9% de la variabilidad en los precios se puede explicar a través de las variables seleccionadas, destacando la importancia de factores como la calidad del inmueble, la ubicación y el estilo del techo en la determinación del valor de las propiedades.

8. Conclusiones.

- ☑ Efectividad del Modelo: Se desarrolló un modelo de regresión avanzada que demostró ser efectivo para predecir los precios de venta de propiedades residenciales en Australia. El modelo lineal alcanzó un R^2 de 0.8894%, indicando que el 88.94% de la variabilidad en los precios puede

ser explicada por las variables seleccionadas, lo que sugiere una buena capacidad predictiva.

☑ Importancia de las Variables: Las variables más influyentes en la predicción del precio de venta fueron identificadas como:
- MSZoning_FV: clasificación de zonificación general de la venta Villa flotante residencial.
- SaleCondition_Alloca: Estado de venta generalmente un condominio con una unidad de garaje.
- SaleType_Otros: Tipo de venta.
- Neighborhood_StoneBr: Vecindario.
- Neighborhood_Veenker Vecindario
- RoofStyle_Gambrel : Tipo del techo
- Functional_Otros
- RoofMatl_Otros Material del techo

Estas variables tienen coeficientes de valores altos en la regresión.

☑ Manejo de Datos Faltantes y Outliers: Se realizó un exhaustivo proceso de limpieza de datos, abordando valores nulos y outliers. Esto mejoró la calidad del conjunto de datos y, por ende, la precisión del modelo. La eliminación o imputación adecuada de datos faltantes fue crucial para evitar sesgos en las predicciones.

☑ Aplicación de Técnicas de Regularización: La implementación de modelos de regresión Lasso y Ridge permitió reducir el riesgo de sobreajuste y mejorar la interpretabilidad del modelo. Lasso, en particular, ayudó a seleccionar características relevantes al eliminar coeficientes no significativos y la eliminación de multicolinealidad.

☑ Optimización Continua: La metodología utilizada para entrenar y validar los modelos sugiere que se pueden realizar mejoras continuas mediante ajustes en los hiperparámetros y exploraciones adicionales de variables. Esto podría resultar en una mayor precisión y robustez del modelo.

☑ Oportunidades Estratégicas: El análisis ha permitido identificar propiedades potencialmente subvaloradas en el mercado australiano, lo que representa oportunidades estratégicas para Surprise Housing al momento de realizar inversiones.

☑ Recomendaciones para Futuras Inversiones: Basado en los resultados obtenidos, se recomienda a Surprise Housing utilizar este modelo predictivo como herramienta clave para tomar decisiones informadas sobre adquisiciones inmobiliarias, maximizando así el retorno sobre la inversión.

En resumen, el proyecto ha proporcionado un marco sólido para la predicción de precios en bienes raíces, permitiendo a Surprise Housing entrar al mercado australiano con una estrategia basada en datos que minimiza riesgos y maximizar oportunidades.

9. Referencias.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). Wiley.

Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models* (3rd ed.). SAGE Publications.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models* (5th ed.). McGraw-Hill/Irwin.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to Linear Regression Analysis* (6th ed.). Wiley.

Schapire, R. E., & Freund, Y. (2012). *Boosting: Foundations and Algorithms*. The MIT Press.

Seber, G. A. F., & Lee, A. J. (2012). *Linear Regression Analysis* (2nd ed.). Wiley.

Tibshirani, R. (1996). *Regression Shrinkage and Selection via the Lasso*. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x