



LABORATORIO 3

Daniel Alejandro Soto Mogollón - Jonnathan Alexander Pérez Ochoa

23 DE JUNIO DE 2024
UNIVERSIDAD ECCI

Caso de estudio

Predicción del Éxito en la Venta de Apartamentos

Objetivo

Determinar las características de los apartamentos que se venden con mayor frecuencia utilizando un modelo basado en datos históricos de FincaRaíz.

Recolección de datos

¿Qué variables (columnas, atributos) de la(s) tabla(s) o base(s) de datos parecen más prometedores?

La información que se encuentra en el archivo trae variables acerca de Apartamentos a la venta, considero que todas las variables son importantes, pero entre ellas resaltan

- Ubicación
- Habitaciones (número habitaciones)
- Nombre (casa/apartamento)
- Parqueaderos
- Área construida
- Área privada
- Estado
- Precio m²
- Antigüedad
- precio

¿Qué variables parecen irrelevantes y pueden ser excluidos?

- Citófono
- Depósito / Bodega
- Barra estilo americano

- Calentador
- Circuito cerrado de TV
- Chimenea

¿Hay suficientes datos para sacar conclusiones generalizables o hacer predicciones precisas?

En la siguiente imagen se puede verificar que se cuentan con distintas variables de importancia para realizar diferentes modelos, dependerá de la predicción que pretendamos dar.

```

Data columns (total 31 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   habitaciones                          8428 non-null   object
1   baños                                8428 non-null   object
2   parqueaderos                          8428 non-null   object
3   area_construida                       8428 non-null   object
4   area_privada                          8428 non-null   object
5   estrato                               8428 non-null   int64
6   estado                                8428 non-null   object
7   antiguedad                            8428 non-null   object
8   administracion                        8428 non-null   object
9   precio_m2                             8428 non-null   object
10  Ascensor                              8428 non-null   int64
11  Circuito cerrado de TV                8428 non-null   int64
12  Parqueadero Visitantes                8428 non-null   int64
13  Portería / Recepción                  8428 non-null   int64
14  Zonas Verdes                          8428 non-null   int64
15  Salón Comunal                         8428 non-null   int64
16  Balcón                                8428 non-null   int64
17  Barra estilo americano                8428 non-null   int64
18  Calentador                            8428 non-null   int64
19  Chimenea                              8428 non-null   int64
20  Citófono                              8428 non-null   int64
21  Cocina Integral                       8428 non-null   int64
22  Terraza                               8428 non-null   int64
23  Vigilancia                            8428 non-null   int64
24  Parques cercanos                     8428 non-null   int64
25  Estudio                              8428 non-null   int64
26  Patio                                 8428 non-null   int64
27  Depósito / Bodega                    8428 non-null   int64
28  nombre                                8428 non-null   object
29  ubicacion                             8428 non-null   object
30  precio                                8428 non-null   object
dtypes: int64(19), object(12)
memory usage: 2.0+ MB

```

¿Hay demasiadas variables para el método de modelado de su elección?

Si, en mi caso utilizaría el modelo de Random Forest debido a su capacidad para manejar tanto variables categóricas como numéricas, su robustez frente a valores faltantes y su interpretabilidad.

¿Está fusionando varias fuentes de datos? Si es así, ¿hay áreas que podrían plantear un problema al fusionar?

Para este caso solo se esta usando una fuente de datos

¿Ha considerado cómo se manejan los valores que faltan en cada uno de sus orígenes de datos?

Si, estos valores faltantes podrían reemplazar por valores que no vayan a ser tenidos en cuenta para la realización de modelos o se podrían clasificar como valores predeterminados para mostrarlos como no obtenidos, sin información, etc.

Describir los datos

¿Cuál es el formato de los datos?

El formato de datos es CSV

¿Cuál es el método utilizado para capturar los datos?

En este caso parece ser el sistema de información o bases de datos de la inmobiliaria.

¿Qué tamaño tiene la base de datos (en número de filas y columnas)?

Tiene una dimensión de 8428 filas por 31 columnas

¿Incluyen los datos una o más variables relevantes para la pregunta de negocio?

Como se evidencia en la imagen hay muchas variables categóricas que clasifican o segmentan gran cantidad de datos en uno solo por lo tanto si se incluyen variables relevantes para el negocio.

¿Qué tipos de datos están presentes (simbólicos, numéricos, etc.)?

Se encuentran los siguientes tipos de datos:

```
habitaciones      object
baños             object
parqueaderos      object
area_construida   object
area_privada      object
estrato           int64
estado            object
antiguedad        object
administracion    object
precio_m2         object
Ascensor          int64
Circuito cerrado de TV  int64
Parqueadero Visitantes int64
Portería / Recepción  int64
Zonas Verdes       int64
Salón Comunal      int64
Balcón            int64
Barra estilo americano int64
Calentador         int64
Chimenea          int64
Citófono          int64
Cocina Integral    int64
Terraza           int64
Vigilancia        int64
Parques cercanos   int64
Estudio           int64
Patio             int64
Depósito / Bodega   int64
nombre            object
ubicacion         object
precio            object
dtype: object
```

¿Ha calculado estadísticas básicas para las variables clave? ¿Qué información le ha proporcionado sobre la cuestión de negocio?

Si, aquí hay algunos

<code>df1['ubicacion'].describe()</code>	<code>df1['habitaciones'].describe()</code>
✓ 0.0s	✓ 0.0s
count 8428	count 8428
unique 1041	unique 17
top Cedritos	top 3
freq 333	freq 4364
Name: ubicacion, dtype: object	Name: habitaciones, dtype: object

<code>df1['nombre'].describe()</code>	<code>df1['parqueaderos'].describe()</code>
✓ 0.0s	✓ 0.0s
count 8428	count 8428
unique 2	unique 12
top Apartamento	top 1
freq 6643	freq 3132
Name: nombre, dtype: object	Name: parqueaderos, dtype: object

<code>df1['area_construida'].describe()</code>	<code>df1['area_privada'].describe()</code>
✓ 0.0s	✓ 0.0s
count 8428	count 8428
unique 696	unique 505
top 60 m ²	top 0 m ²
freq 135	freq 2082
Name: area_construida, dtype: object	Name: area_privada, dtype: object

<code>df1['estado'].describe()</code>	<code>df1['estrato'].describe()</code>
✓ 0.0s	✓ 0.0s
count 8428	count 8428.000000
unique 4	mean 4.296749
top No definida	std 1.263955
freq 3838	min 0.000000
Name: estado, dtype: object	25% 3.000000
	50% 4.000000
	75% 5.000000
	max 6.000000
	Name: estrato, dtype: float64

<code>df1['precio'].describe()</code>	<code>df1['precio_m2'].describe()</code>
✓ 0.0s	✓ 0.0s
count 8428	count 8428
unique 919	unique 3854
top 350000000	top \$ 5.000.000*m ²
freq 126	freq 122
Name: precio, dtype: object	Name: precio_m2, dtype: object

```
df1['antigüedad'].describe()
✓ 0.0s
```

count	8428
unique	6
top	16 a 30 años
freq	2626
Name: antigüedad, dtype: object	

Exploración de datos

¿Qué tipo de hipótesis se ha formado sobre los datos?

Una de las hipótesis podría plantearse sería algo como que características de casas o apartamentos son más ofertadas y que tipos de personas de diferentes estratos compran estas casas

¿Qué variables parecen prometedoras para un análisis más profundo?

- Ubicación
- Área construida
- Estado
- Precio m²
- Antigüedad
- Precio

¿Sus exploraciones han revelado nuevas características sobre los datos?

De momento se ha realizado revisión de valores únicos de variables, verificación de tipos de datos, etc. no se han detectado nuevas características.

¿Cómo han cambiado estas exploraciones su hipótesis inicial?

El objetivo no ha cambiado.

¿Considera que debería reformular el alcance del proyecto?

No

¿Puede identificar subconjuntos particulares de datos para su uso posterior?

Se podrían dividir entre apartamentos y casas, para tener una mayor focalización de la población

Verificar la calidad de datos

Identificar datos faltantes

¿Ha identificado variables faltantes y campos en blanco? Si es así, ¿Hay algún significado detrás de tales valores faltantes?

```
Data columns (total 31 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   habitaciones                          8428 non-null   object
1   baños                                 8428 non-null   object
2   parqueaderos                          8428 non-null   object
3   area_construida                       8428 non-null   object
4   area_privada                          8428 non-null   object
5   estrato                               8428 non-null   int64
6   estado                                8428 non-null   object
7   antigüedad                            8428 non-null   object
8   administracion                        8428 non-null   object
9   precio_m2                             8428 non-null   object
10  Ascensor                              8428 non-null   int64
11  Circuito cerrado de TV                 8428 non-null   int64
12  Parqueadero Visitantes                 8428 non-null   int64
13  Portería / Recepción                   8428 non-null   int64
14  Zonas Verdes                           8428 non-null   int64
15  Salón Comunal                          8428 non-null   int64
16  Balcón                                  8428 non-null   int64
17  Barra estilo americano                  8428 non-null   int64
18  Calentador                             8428 non-null   int64
19  Chimenea                               8428 non-null   int64
20  Citófono                               8428 non-null   int64
21  Cocina Integral                         8428 non-null   int64
22  Terraza                                8428 non-null   int64
23  Vigilancia                             8428 non-null   int64
24  Parques cercanos                       8428 non-null   int64
25  Estudio                                8428 non-null   int64
26  Patio                                  8428 non-null   int64
27  Depósito / Bodega                      8428 non-null   int64
28  nombre                                 8428 non-null   object
29  ubicacion                              8428 non-null   object
30  precio                                 8428 non-null   object
dtypes: int64(19), object(12)
memory usage: 2.0+ MB
```

¿Hay inconsistencias ortográficas que puedan causar problemas en fusiones o transformaciones posteriores?

Si, se realizó exploración de las variables y en algunas se encontró diferencia entre mayúsculas y minúsculas, espacio, tildes, etc.

También tenemos que variables como el precio son de tipo String por razón de que tiene el símbolo pesos, esto impide realizar cálculos.

Ejemplo:

```
for i in df1['ubicacion'].unique():  
    print (i)  
✓ 0.0s
```

La esperanza
La florida occidental
LA MANUELITA
Chico Reservado
MURILLO TORO QUIROGA
Santa teresita
Ub. santa luisa
san simon
San ignacio
Los libertadores san luis
Las aguas
La soledad
La Veracruz
PARCELAS
San patricio
Santa ana occidental
Super manzana 8
Plenitud
Colina y Alrededores
Sinai
La paz
San vicente ferrer
El Moral
Santiago
VILLA GLADYS
Plaza de las Américas
—

¿Ha explorado las desviaciones para determinar si son "ruido" o fenómenos que vale la pena analizar más a fondo?

¿Ha realizado una comprobación de plausibilidad de los valores? Tome notas sobre cualquier conflicto aparente (como adolescentes con altos niveles de ingresos).

Si, en particular tenemos un caso en el que al parecer el precio no concuerda con las características normales de casas y apartamentos

```
df1.max()
[59] ✓ 0.0s

...   ubicacion      zuñiga
      habitaciones  No definida
      nombre        Casa
      parqueaderos  Más de 10
      area_construida  99,7 m²
      area_privada    99,41 m²
      estado        Remodelado
      estrato        6
      precio         999000000
      antiguedad      más de 30 años
      precio_m2       $ 969.305,33*m²
      dtype: object

df1.min()
[60] ✓ 0.0s

...   ubicacion      12 de Octubre
      habitaciones      1
      nombre        Apartamento
      parqueaderos      0
      area_construida    1 m²
      area_privada       0 m²
      estado          Bueno
      estrato          0
      precio         1000000000
      antiguedad      1 a 8 años
      precio_m2       $ 1.028.571,43*m²
      dtype: object
```

¿Ha considerado excluir datos que no tienen impacto en sus hipótesis?

Si, existen variables que no tienen tanta importancia para el objetivo del caso de estudio como los son:

- Citófono
- Barra estilo americano
- Calentador
- Circuito cerrado de TV
- Chimenea

¿Los datos se almacenan en archivos planos? Si es así, ¿Son los delimitadores

coherentes entre los archivos?

El archivo del caso de estudio es un archivo plano, el delimitador no es coherente ya que hay un valor en la columna precio que contiene coma, por lo cual al separarlo causa conflicto dejando algunas variables nulas.

¿Cada registro contiene el mismo número de campos?

Se puede evidenciar que ningún registro es nulo

```
df.isna().sum()
✓ 0.0s
```

habitaciones	0
los	0
ns_	0
parqueaderos	0
area_construida	0
area_privada	0
estrato	0
estado	0
antiguedad	0
administracion	0
precio_m2	0
Ascensor	0
Círculo cerrado de TV	0
Parqueadero Visitantes	0
Portería / Recepción	0
Zonas Verdes	0
Salón Comunal	0
Balcón	0
Barra estilo americano	0
Calentador	0
Chimenea	0
Citófono	0
Cocina Integral	0
Terraza	0
Vigilancia	0
Parques cercanos	0
Estudio	0
Patio	0
Depósito / Bodega	0
nombre	0
ubicacion	0