

Segunda entrega de proyecto

POR:

Luisa María Castro Cortez

MATERIA:

Introducción a la inteligencia artificial

PROFESOR:

Raul Ramos Pollan



UNIVERSIDAD DE ANTIOQUIA FACULTAD DE INGENIERÍA

MEDELLÍN 2023

Contenido

1. Planteamiento del problema	3
1.1 Dataset	3
1.2 Métricas	6
2 Exploración de variables	7
2.1 Análisis de la variable objetivo SalePrice	7
2.2 Descubrimiento de los tipos de datos	8
3 Bibliografía	14

1. Planteamiento del problema

Si nos ponemos en la tarea de preguntar a un comprador de vivienda que nos describa la casa de sus sueños probablemente no empezará por la altura del techo o la proximidad a una estación del metro. Pero existen muchas variables que influyen en las negociaciones sobre el precio de venta de una propiedad. Es por esto por lo que se desea desarrollar un modelo que permita estimar el precio de venta de las viviendas, basado en las variables que pueden presentar las viviendas, al tener mejores estimaciones, se podrá generar un mayor interés por parte de los compradores de vivienda al momento de buscar una nueva propiedad.

1.1 Dataset

El dataset a utilizar proviene de una competencia de kaggle en la cual se proporcionan datos con 79 variables explicativas que describen aspectos de las viviendas residenciales en Ames, Iowa. El dataset este compuesto por un conjunto de archivos .csv y .txt que proporcionan la información requerida.

El archivo que contiene los datos de las viviendas es nombrado data_description y contiene la siguiente información:

- SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of property
- LandContour: Flatness of the property
- Utilities: Type of utilities available
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city limits

- Condition1: Proximity to main road or railroad
- Condition2: Proximity to main road or railroad (if a second is present)
- BldgType: Type of dwelling
- HouseStyle: Style of dwelling
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- RoofStyle: Type of roof
- RoofMatl: Roof material
- Exterior1st: Exterior covering on house
- Exterior2nd: Exterior covering on house (if more than one material)
- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area in square feet
- ExterQual: Exterior material quality
- ExterCond: Present condition of the material on the exterior
- Foundation: Type of foundation
- BsmtQual: Height of the basement
- BsmtCond: General condition of the basement
- BsmtExposure: Walkout or garden level basement walls
- BsmtFinType1: Quality of basement finished area
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinType2: Quality of second finished area (if present)
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- Heating: Type of heating
- HeatingQC: Heating quality and condition
- CentralAir: Central air conditioning
- Electrical: Electrical system
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens

- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Functional: Home functionality rating
- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- GarageType: Garage location
- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- GarageQual: Garage quality
- GarageCond: Garage condition
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet
- 3SsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet
- PoolArea: Pool area in square feet
- PoolQC: Pool quality
- Fence: Fence quality
- MiscFeature: Miscellaneous feature not covered in other categories
- MiscVal: \$Value of miscellaneous feature
- MoSold: Month Sold
- YrSold: Year Sold
- SaleType: Type of sale
- SaleCondition: Condition of sale

train.csv - el conjunto de entrenamiento

test.csv - el conjunto de prueba

sample_submission.csv - un envío de referencia de una regresión lineal sobre el año y el mes de la venta, los metros cuadrados del lote y el número de dormitorios.

1.2 Métricas

La métrica de evaluación principal para el modelo será error cuadrático medio (RMSE) entre el logaritmo del valor predicho y el logaritmo del precio de venta observado. (Tomar los logaritmos significa que los errores en la predicción de casas caras y casas baratas afectarán por igual al resultado). el cual se calcula mediante la siguiente expresión:

$$\epsilon = \sqrt{\frac{\sum_{i=1}^n (P_i - O_i)^2}{n}}$$

Donde ϵ es el valor del RMSE, n es el número total de observaciones en el dataset, p_i es la predicción de la variable objetivo y O_i es el valor real de la variable objetivo.

Por otra parte, en cuanto a la métrica de negocio, se tiene interés en que las predicciones sean lo suficientemente confiables para saber precio de venta de las viviendas. Con esta información se pueden realizar análisis financieros para determinar qué tan viables pueden ser comprar ciertas viviendas en particular.

2. Exploración de variables

2.1 Análisis de la variable objetivo SalePrice

Para empezar con la exploración de variables, se analiza el comportamiento de la distribución que tiene la variable objetivo SalePrice, dicho comportamiento se muestra en la Figura 1, donde se puede observar que la variable objetivo tiene una asimetría hacia la derecha. Por lo que se aplica una transformación logarítmica, cuyo resultado se muestra en la Figura 2.

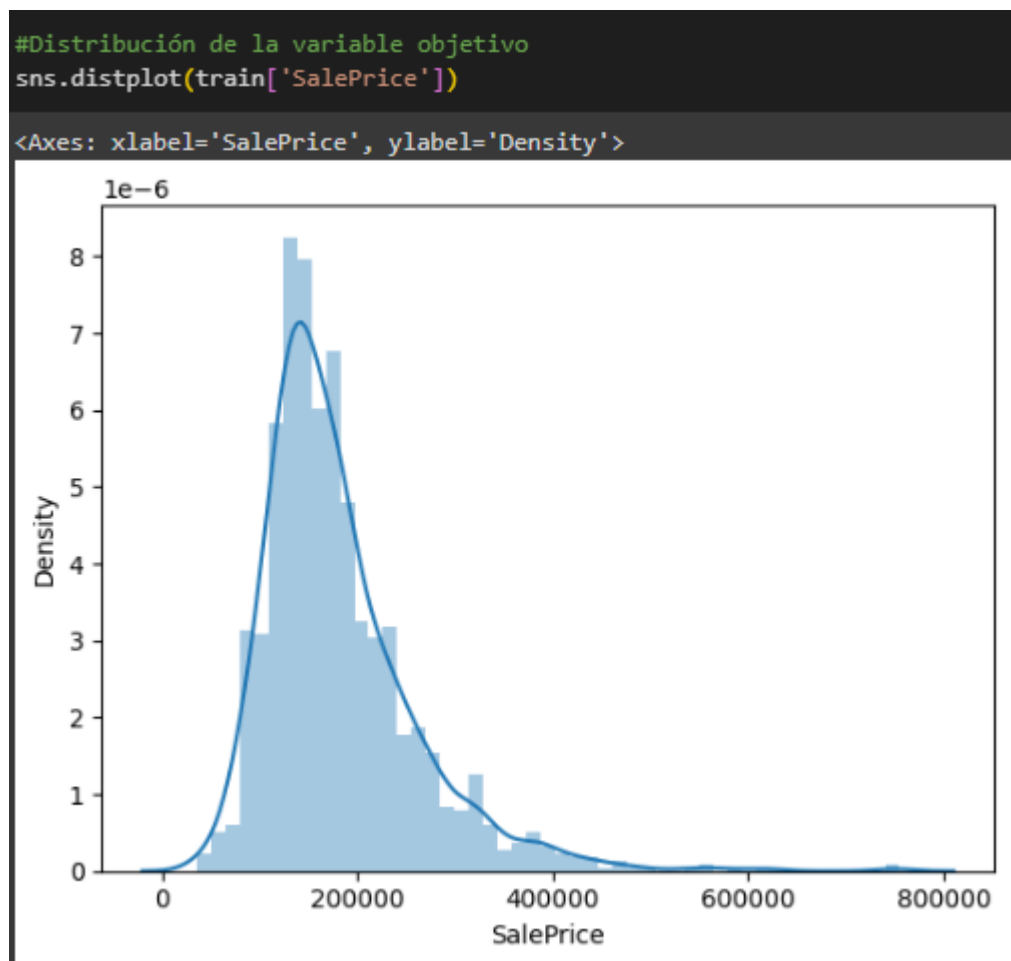


Figura 1. Distribución de la variable objetivo.

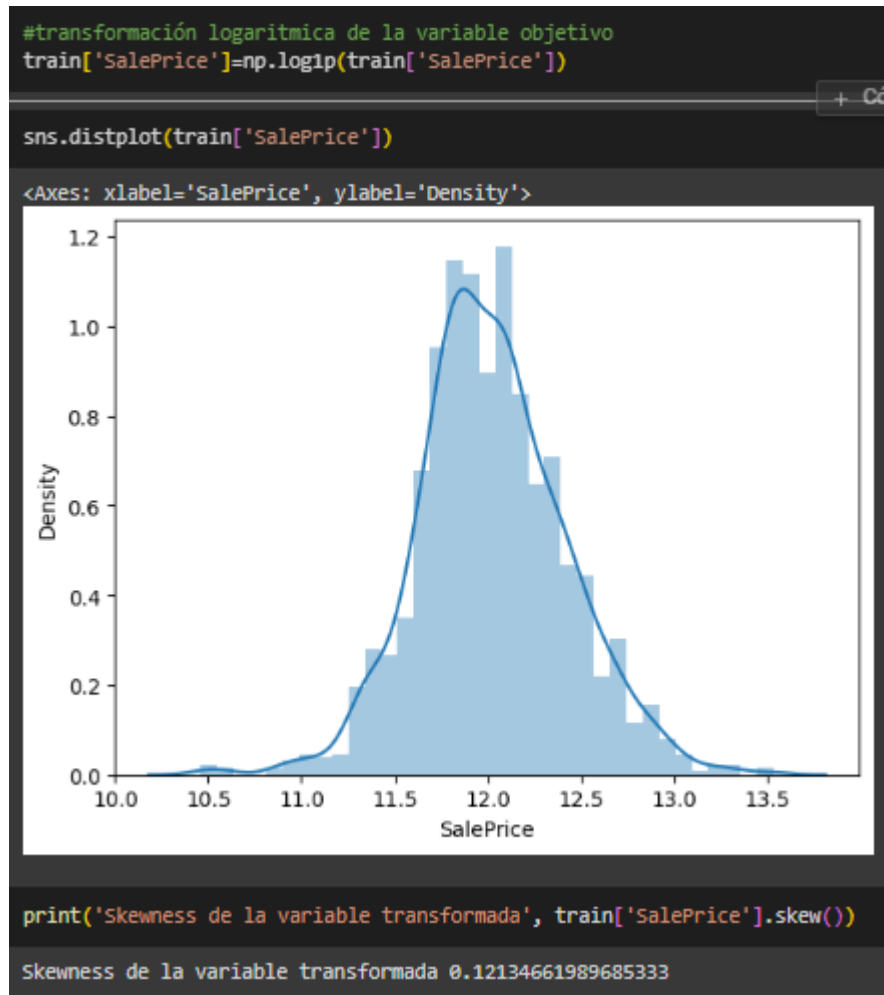


Figura 2. Distribución de la variable objetivo después de la transformación logarítmica.

En la Figura 2, se puede observar que la distribución de la variable objetivo luego de la transformación logarítmica ya tiene un comportamiento más adecuado para realizar un análisis, y es esta variable transformada la que se usará en el entrenamiento y pruebas de los algoritmos. Cabe destacar que la asimetría o skewness de la variable objetivo antes de la transformación era de 1.88, y el valor luego de la transformación es de 0.12, lo cual refleja una gran reducción.

2.2 Descubrimiento de los tipos de datos

```
train.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 81 columns):
```


#	COLUMN	NON-NULL	COUNT	DTYPE
0	Id	1460	non-null	int64
1	MSSubClass	1460	non-null	int64
2	MSZoning	1460	non-null	object
3	LotFrontage	1201	non-null	float64
4	LotArea	1460	non-null	int64
5	Street	1460	non-null	object
6	Alley	91	non-null	object
7	LotShape	1460	non-null	object
8	LandContour	1460	non-null	object
9	Utilities	1460	non-null	object
10	LotConfig	1460	non-null	object
11	LandSlope	1460	non-null	object
12	Neighborhood	1460	non-null	object
13	Condition1	1460	non-null	object
14	Condition2	1460	non-null	object
15	BldgType	1460	non-null	object
16	HouseStyle	1460	non-null	object
17	OverallQual	1460	non-null	int64
18	OverallCond	1460	non-null	int64
19	YearBuilt	1460	non-null	int64
20	YearRemodAdd	1460	non-null	int64
21	RoofStyle	1460	non-null	object
22	RoofMatl	1460	non-null	object
23	Exterior1st	1460	non-null	object
24	Exterior2nd	1460	non-null	object
25	MasVnrType	1452	non-null	object
26	MasVnrArea	1452	non-null	float64
27	ExterQual	1460	non-null	object
28	ExterCond	1460	non-null	object
29	Foundation	1460	non-null	object
30	BsmtQual	1423	non-null	object
31	BsmtCond	1423	non-null	object
32	BsmtExposure	1422	non-null	object
33	BsmtFinType1	1423	non-null	object
34	BsmtFinSF1	1460	non-null	int64
35	BsmtFinType2	1422	non-null	object
36	BsmtFinSF2	1460	non-null	int64
37	BsmtUnfSF	1460	non-null	int64
38	TotalBsmtSF	1460	non-null	int64
39	Heating	1460	non-null	object

40	HeatingQC	1460	non-null	object
41	CentralAir	1460	non-null	object
42	Electrical	1459	non-null	object
43	1stFlrSF	1460	non-null	int64
44	2ndFlrSF	1460	non-null	int64
45	LowQualFinSF	1460	non-null	int64
46	GrLivArea	1460	non-null	int64
47	BsmtFullBath	1460	non-null	int64
48	BsmtHalfBath	1460	non-null	int64
49	FullBath	1460	non-null	int64
50	HalfBath	1460	non-null	int64
51	BedroomAbvGr	1460	non-null	int64
52	KitchenAbvGr	1460	non-null	int64
53	KitchenQual	1460	non-null	object
54	TotRmsAbvGrd	1460	non-null	int64
55	Functional	1460	non-null	object
56	Fireplaces	1460	non-null	int64
57	FireplaceQu	770	non-null	object
58	GarageType	1379	non-null	object
59	GarageYrBlt	1379	non-null	float64
60	GarageFinish	1379	non-null	object
61	GarageCars	1460	non-null	int64
62	GarageArea	1460	non-null	int64
63	GarageQual	1379	non-null	object
64	GarageCond	1379	non-null	object
65	PavedDrive	1460	non-null	object
66	WoodDeckSF	1460	non-null	int64
67	OpenPorchSF	1460	non-null	int64
68	EnclosedPorch	1460	non-null	int64
69	3SsnPorch	1460	non-null	int64
70	ScreenPorch	1460	non-null	int64
71	PoolArea	1460	non-null	int64
72	PoolQC	7	non-null	object
73	Fence	281	non-null	object
74	MiscFeature	54	non-null	object
75	MiscVal	1460	non-null	int64
76	MoSold	1460	non-null	int64
77	YrSold	1460	non-null	int64
78	SaleType	1460	non-null	object
79	SaleCondition	1460	non-null	object
80	SalePrice	1460	non-null	float64

Tabla 1 Información de las variables

De las 79 variables que tenemos en cuenta para realizar la predicción del SalePrice el 48% son numéricas y el 52% son categóricas las cuales tendremos que transformar en el proceso de tratamiento de datos. De las 79 variables, 19 variables tienen valores nulos, los cuales tendremos que limpiar antes de aplicar los modelos.

2.3 Correlación de variables

La Tabla 2 muestra los valores de correlación que existen entre las diferentes variables con la variable objetivo. Se puede observar que OverallQual, GrLivArea, GarageCars, GarageArea son las variables que tienen la mayor correlación. También se puede observar que las variables iscVal, OverallCond, YrSold, LowQualFinSF, MSSubClass, KitchenAbvGr, EnclosedPorch, tienen una correlación tan baja que puede decirse que no están relacionadas con la variable objetivo.

	SALEPRICE
OVERALLQUAL	0.81718461
GRLIVAREA	0.70092699
GARAGECARS	0.68062487
GARAGEAREA	0.65088768
TOTALBSMTSF	0.61213423
1STFLRSF	0.59698132
FULLBATH	0.59477066
YEARBUILT	0.58657019
YEARREMODADD	0.56560778
GARAGEYRBLT	0.54107278
TOTRMSABVGRD	0.5344224
FIREPLACES	0.48944955
MASVNRAREA	0.43080896
BSMTFINSF1	0.37202325
LOTFRONTAGE	0.35587862
WOODDECKSF	0.33413517
OPENPORCHSF	0.32105325
2NDFLRSF	0.31930014
HALFBATH	0.31398222
LOTAREA	0.25732007
BSMTFULLBATH	0.23622416
BSMTUNFSF	0.22198516
BEDROOMABVGR	0.20904343
SCREENPORCH	0.12120759
POOLAREA	0.0697979
MOSOLD	0.0573295

3SSNPORCH	0.0549002
BSMTFINSF2	0.00483229
BSMTHALFBATH	-0.00514924
MISCVAL	-0.02002082
OVERALLCOND	-0.03686845
YRSOLD	-0.03726291
LOWQUALFINSF	-0.03796279
MSSUBCLASS	-0.07395917
KITCHENABVGR	-0.14754816
ENCLOSEDPORCH	-0.14905023

Tabla 2 Correlación de variables con la variable objetivo

Histograms For the Different Features

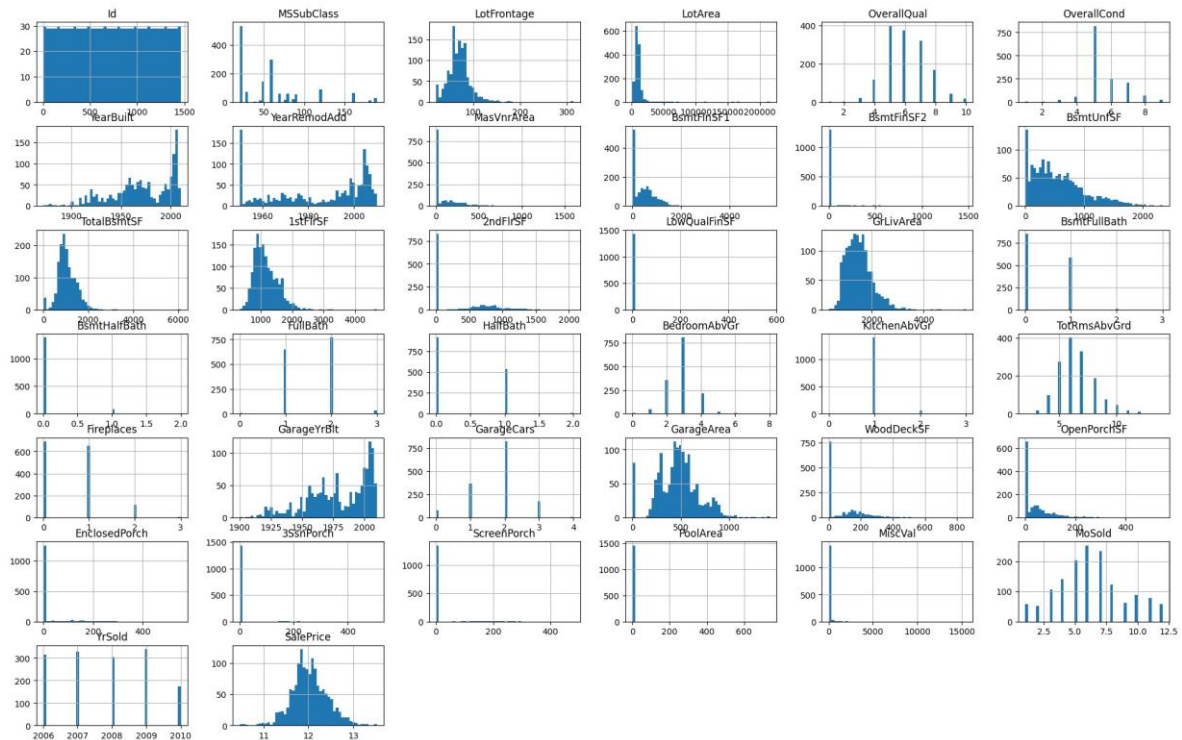


Figura 4 Distribución de variables numéricas

3. Tratamiento de datos

3.1 Eliminación de las columnas con muchos datos faltantes

Como se mostró en la sección anterior, existen variables que tienen datos faltantes. Para este caso se considerará que una variable que tenga más del 50% de datos faltantes será eliminada ya que no aportará la suficiente información al modelo. De esta manera las variables que son eliminadas del dataset son Alley, PoolQC, Fence, MiscFeature.

```

criterio = len(train) * 0.5 #criterio para eliminar la columna (50% de las filas que se tienen)
train.dropna(axis=1, thresh = criterio, inplace = True) #eliminación de las columnas con 50% o más de datos faltantes
print('New Shape of Train Data:',train.shape)

New Shape of Train Data: (1460, 77)

2. Relleno de datos numéricos faltantes

train.info()

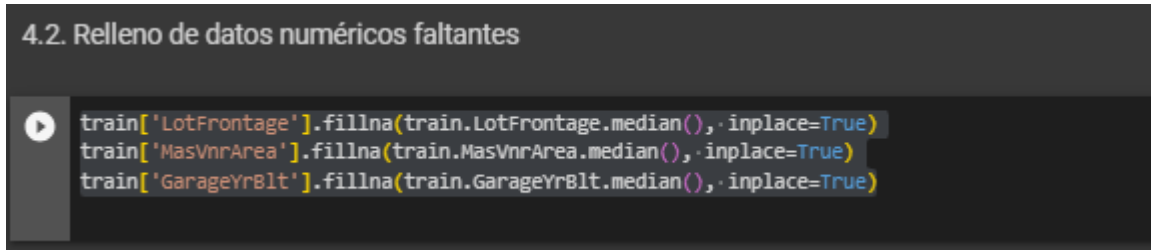
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 77 columns):

```

Figura 5 Eliminación de variables con datos nulos superiores al 50%

3.2 Relleno de datos faltantes

Para el resto de las variables numéricas que contienen datos faltantes que no son eliminadas rellenan dichos valores nulos con la media de sus datos. La Figura 6 muestra el código usado para rellenar dichos valores.



```
4.2. Relleno de datos numéricos faltantes

train['LotFrontage'].fillna(train.LotFrontage.median(), inplace=True)
train['MasVnrArea'].fillna(train.MasVnrArea.median(), inplace=True)
train['GarageYrBlt'].fillna(train.GarageYrBlt.median(), inplace=True)
```

Figura 6 Relleno de datos numéricos faltantes

3.3 Transformación de variables categóricas

Las variables categóricas no pueden ser usadas en el entrenamiento de un algoritmo, por lo que se debe hacer una transformación para convertirlas a variables numéricas. Aún me encuentro realizando este proceso ya que son más de 30 variables y es un proceso extenso.

4. Bibliografía

- Anna Montoya, DataCanary. (2016). House Prices - Advanced Regression Techniques. Kaggle. <https://kaggle.com/competitions/house-prices-advancedregression-techniques>