SEGUNDA ENTREGA DEL PROYECTO

POR:
Juan David Sandoval Guerrero
Daniela Tuberquia Villa

MATERIA:
Introducción a la inteligencia artificial.

PROFESOR:
Raul Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA FACULTAD DE INGENIERÍA 2022.

1. **Planteamiento del problema.**
   El modelo que se desarrolla basado en la competencia que tiene como título: “House price”, el cual dependiendo de las características de una vivienda tales como su superficie, localización, entre otros. Se quiere llegar a predecir el valor final de la vivienda en el mercado.

2. **Dataset.**
   El dataset a utilizar se puede encontrar en el siguiente link:https://www.kaggle.com/competitions/house-prices-advanced-regression-techniq ues/data.
   El cual consta de 1460 muestras (casas) y de 80 columnas las cuales describen casi todas las características de las viviendas.

3. **Métrica.**
   La métrica de evaluación principal que se utilizara en este proyecto será el error cuadrático medio (RMSE), el cual consiste es la diferencia entre los valores predichos por un modelo y los valores observados.

$$\sqrt{\frac{1}{N}\sum (yi - \hat{y}i)}$$

   donde N es el número total de datos del dataset, yi es el valor real y $\hat{y}i$ Es el valor de la predicción.
   En la métrica de negocio, se desea que las predicciones sean lo suficientemente confiables, para saber el valor de las viviendas y así hacer un análisis más detallado en este negocio inmobiliario.

4. **Desempeño.**
   Lo que se desea de este modelo es predecir los precios de las ventas de las casas, según los requerimientos de cada cliente para así darle un precio estimado de la vivienda en el primer momento en el cual el cliente consulta en la empresa por su vivienda, el modelo no sería fiable si tiene un error mayor al 25% ya que se perdería credibilidad.

5. **Procesamiento de datos.**
   Para iniciar la exploración de los datos a ser tratados, se usa el archivo “Train”, el cual contiene 81 variables y 1460 muestras, lo primero que se realizar es clasificarla en datos numéricos y datos categóricos, los datos numéricos son subdivididos en: temporales, continuos y discretos, mientras que los datos categóricos son solo de un tipo. A continuación, se muestra el filtrado de los datos mencionados anteriormente:
   5.1 Filtrado de datos numéricos:
   Para lograr obtener los datos numéricos, se filtra el dataframe inicial, para

este caso se obtienen 38 variables numéricas, las cuales serán subdivididas en las categorías mencionadas anteriormente, los datos numéricos obtenidos son los siguientes:

MSSubClass, LotFrontage, LotArea, OverallQual, OverallCond, YearBuilt, YearRemodAdd, MasVnrArea, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, 1stFlrSF, 2ndFlrSF, LowQualFinSF, GrLivArea, BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd, Fireplaces, GarageYrBlt, GarageCars, GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, 3SsnPorch, ScreenPorch, PoolArea, MiscValMoSold, YrSold, SalePrice.

Luego de obtener los datos numéricos, se filtran en las tres categorías numéricas.

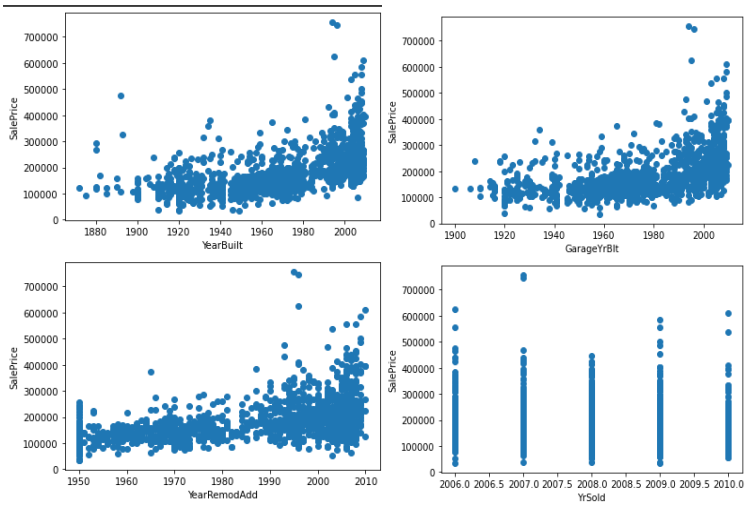
5.2 Filtrado de datos numéricos temporales

En los datos numéricos temporales se obtienen las variables de tipo temporal, los cuales son: año de construcción, año de remodelación, año de remodelación del garaje y el año de puesta a tierra. A continuación, un resumen de las variables temporales:

	count	mean	std	min	25%	50%	75%	max
YearBuilt	1460.0	1971.267808	30.202904	1872.0	1954.0	1973.0	2000.0	2010.0
YearRemodAdd	1460.0	1984.865753	20.645407	1950.0	1967.0	1994.0	2004.0	2010.0
GarageYrBlt	1379.0	1978.506164	24.689725	1900.0	1961.0	1980.0	2002.0	2010.0
YrSold	1460.0	2007.815753	1.328095	2006.0	2007.0	2008.0	2009.0	2010.0

Tabla 1. Resumen de variables temporales.

Gráfica de los datos temporales vs precio de venta:



De las anteriores gráficas se puede concluir que a mayor sea el año de construcción de la vivienda, el año de remodelación o el año de construcción del garaje, mayor es el precio de venta, mientras que para el año de puesta a la venta no se ve ninguna relación especifica, más adelante se busca el nivel de correlación entre estas variables y el precio de venta.

5.3 Filtrado de datos numéricos discretos

Para este caso se realizó un filtrado para los datos numéricos de tipo discretos, obteniendo las siguientes 17 variables:

	count	mean	std	min	25%	50%	75%	max
MSSubClass	1460.0	56.897260	42.300571	20.0	20.0	50.0	70.0	190.0
OverallQual	1460.0	6.099315	1.382997	1.0	5.0	6.0	7.0	10.0
OverallCond	1460.0	5.575342	1.112799	1.0	5.0	5.0	6.0	9.0
LowQualFinSF	1460.0	5.844521	48.623081	0.0	0.0	0.0	0.0	572.0
BsmtFullBath	1460.0	0.425342	0.518911	0.0	0.0	0.0	1.0	3.0
BsmtHalfBath	1460.0	0.057534	0.238753	0.0	0.0	0.0	0.0	2.0
FullBath	1460.0	1.565068	0.550916	0.0	1.0	2.0	2.0	3.0
HalfBath	1460.0	0.382877	0.502885	0.0	0.0	0.0	1.0	2.0
BedroomAbvGr	1460.0	2.866438	0.815778	0.0	2.0	3.0	3.0	8.0
KitchenAbvGr	1460.0	1.046575	0.220338	0.0	1.0	1.0	1.0	3.0
TotRmsAbvGrd	1460.0	6.517808	1.625393	2.0	5.0	6.0	7.0	14.0
Fireplaces	1460.0	0.613014	0.644666	0.0	0.0	1.0	1.0	3.0
GarageCars	1460.0	1.767123	0.747315	0.0	1.0	2.0	2.0	4.0
3SsnPorch	1460.0	3.409589	29.317331	0.0	0.0	0.0	0.0	508.0
PoolArea	1460.0	2.758904	40.177307	0.0	0.0	0.0	0.0	738.0
MiscVal	1460.0	43.489041	496.123024	0.0	0.0	0.0	0.0	15500.0
MoSold	1460.0	6.321918	2.703626	1.0	5.0	6.0	8.0	12.0

Tabla 2. Resumen de datos numéricos discretos.

5.4 Filtrado de datos numéricos continuos

Para este caso se realizó un filtrado para los datos numéricos de tipo discretos, obteniendo las siguientes 16 variables:

	count	mean	std	min	25%	50%	75%	max
LotFrontage	1201.0	70.049958	24.284752	21.0	59.00	69.0	80.00	313.0
LotArea	1460.0	10516.828082	9981.264932	1300.0	7553.50	9478.5	11601.50	215245.0
MasVnrArea	1452.0	103.685262	181.066207	0.0	0.00	0.0	166.00	1600.0
BsmtFinSF1	1460.0	443.639726	456.098091	0.0	0.00	383.5	712.25	5644.0
BsmtFinSF2	1460.0	46.549315	161.319273	0.0	0.00	0.0	0.00	1474.0
BsmtUnfSF	1460.0	567.240411	441.866955	0.0	223.00	477.5	808.00	2336.0
TotalBsmtSF	1460.0	1057.429452	438.705324	0.0	795.75	991.5	1298.25	6110.0
1stFlrSF	1460.0	1162.626712	386.587738	334.0	882.00	1087.0	1391.25	4692.0
2ndFlrSF	1460.0	346.992466	436.528436	0.0	0.00	0.0	728.00	2065.0
GrLivArea	1460.0	1515.463699	525.480383	334.0	1129.50	1464.0	1776.75	5642.0
GarageArea	1460.0	472.980137	213.804841	0.0	334.50	480.0	576.00	1418.0
WoodDeckSF	1460.0	94.244521	125.338794	0.0	0.00	0.0	168.00	857.0
OpenPorchSF	1460.0	46.660274	66.256028	0.0	0.00	25.0	68.00	547.0
EnclosedPorch	1460.0	21.954110	61.119149	0.0	0.00	0.0	0.00	552.0
ScreenPorch	1460.0	15.060959	55.757415	0.0	0.00	0.0	0.00	480.0
SalePrice	1460.0	180921.195890	79442.502883	34900.0	129975.00	163000.0	214000.00	755000.0

Tabla 3. Resumen de datos numéricos continuos.

5.5 Correlación de datos numéricos

Después de realizar el filtrado de las variables numéricas y como se había mencionado anteriormente, se procede a realizar un filtrado de correlación con respecto al precio de venta, todo esto haciendo uso del coeficiente de correlación de Pearson, para este caso, se seleccionan las variables con correlación mayor a 0.2, obteniendo las siguientes variables:

LotFrontage,LotArea,OverallQual,YearBuilt,YearRemodAdd,MasVnrArea, BsmtFinSF1,BsmtUnfSF,TotalBsmtSF,1stFlrSF,2ndFlrSF,GrLivArea,Bsmt FullBath,FullBath,HalfBath,TotRmsAbvGrd,Fireplaces,GarageYrBlt,Garag eCars,GarageArea,WoodDeckSF,OpenPorchSF,SalePrice.

6. Filtrado de datos categóricos

Se procede a filtrar los datos de tipo categóricos del dataframe inicial, del cual se obtienen 43 variables de tipo categóricas, dentro de estos datos categóricos, se encuentran variables únicas, que nos indican la categoría a la cual pertenece. En este proceso a cada variable única se le asigna un número dentro de su tipo de dato categórico, como se muestra a continuación:

	MSZoning	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope
0	3	1	2	3	3	0	4	0
1	3	1	2	3	3	0	2	0
2	3	1	2	0	3	0	4	0
3	3	1	2	0	3	0	0	0
4	3	1	2	0	3	0	2	0
...	...	...	...	...	...	...	...	...
1455	3	1	2	3	3	0	4	0
1456	3	1	2	3	3	0	4	0
1457	3	1	2	3	3	0	4	0
1458	3	1	2	3	3	0	4	0
1459	3	1	2	3	3	0	4	0

Tabla 4. filtrado de datos categóricos

6.1 Correlación de datos categóricos

Aplicando para este caso el coeficiente de correlación de Pearson, las variables con una correlación mayor a 0.2 son las siguientes:

Neighborhood,RoofStyle,Foundation,CentralAir,Electrical,PavedDrive,Sale Condition,SalePrice.

Conclusión

Después de realizar los filtrados correspondientes a los datos numéricos y a los datos categóricos, se obtuvieron 7 variables de interés para los datos categóricos y para los datos numéricos se obtuvieron 22 variables de interés, para un total de 29 variables de interés respecto a nuestra variable objetivo que es el precio de venta.