Análisis de los datos

Introducción

En el presente documento se buscará analizar el dataset a utilizar en la búsqueda de poder predecir el precio de venta de un inmueble.

Pasos a seguir a la hora de analizar la data:

1. Comprender el problema analizando cada una de las variables, determinando su significado y su importancia dentro del problema.
2. Estudiar la variable a predecir (variable dependiente).
3. Estudiar relación entre las diversas variables dependientes y la variable a predecir.
4. Tratamiento de la data, pasos necesarios antes de aplicar el modelo de aprendizaje automático elegido.

Análisis de los datos

Dataset a utilizar

Dataset House prices: Advanced Regression Techniques

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

Cantidad de observaciones: **1406**

Cantidad de variables numéricas: **36**

Cantidad de variables categóricas: **43**

Variables del dataset

SalePrice: Precio de la propiedad, en dólares. *Numérico*.

MSSubClass: Tipo de edificio. *Numérico*.

MSZoning: Ubicación (zona) general del inmueble. *Polinomial.*

LotFrontpage: Distancia entre la calle y la propiedad (en pies). *Numérico.*

LotArea: Tamaño de la propiedad (en pies cuadrados). *Numérico.*

Street: Tipo de calle. *Binomial.*

Alley: Tipo de callejón (calle lateral). *Polinomial.*

LotShape: Forma general de la propiedad. *Polinomial.*

LandContour: Planitud de la propiedad. *Polinomial.*

Utilities: Tipo de utilidades disponibles. *Polinomial.*

LotConfig: Configuración del lote. *Polinomial.*

LandSlope: Planitud de la propiedad. *Polinomial.*

Neighborhood: Barrio dentro de la ciudad de Ames. *Polinomial.*

Condition1: Proximidad a varias condiciones. *Polinomial.*

Condition2: Proximidad a varias condiciones (si hay más de una presente). *Polinomial.*

BldgType: Tipo de vivienda. *Polinomial.*

HouseStyle: Estilo de vivienda. *Polinomial.*

OverallQual: Calidad del material utilizado en general y en el acabado de la propiedad. *Numérico.*

OverallCond: Calificación general de la propiedad. *Numérico.*

YearBuilt: Año de construcción. *Numérico.*

YearRemodAdd: Año de remodelación (misma que de construcción si no ha tenido remodelaciones). *Numérico.*

RoofStyle: Tipo de techo. *Polinomial.*

RoofMatl: Material del techo. *Numérico.*

Exterior1st: Construcción perimetral de la casa (cubierta exterior, material). *Polinomial.*

Exterior2nd: Construcción perimetral de la casa (si hay más de un material). *Polinomial.*

MasVnrType: Tipo de mampostería. *Polinomial.*

MasVnrArea: Área de mampostería (en pie cuadrado). *Numérico.*

ExterQual: Calidad del material exterior de la propiedad. *Polinomial.*

ExterCond: Estado actual del material en el exterior. *Polinomial.*

Foundation: Tipo de cimientos. *Polinomial.*

BsmtQual: Evaluación de la altura del sótano. *Polinomial.*

BsmtCond: Evaluación general del sótano. *Polinomial.*

BsmtExposure: Nivel de exposición en paredes a nivel de jardín o al exterior. *Polinomial.*

BsmtFinType1: Calidad en terminado del sótano. *Polinomial.*

BsmtFinSF1: Área (pies cuadrados) asociado a variable BsmtFinType1. *Numérico.*

BsmtFinType2: Calidad en terminado del sótano (otra opción si es necesaria). *Polinomial.*

BsmtFinSF2: Área (pies cuadrados) asociado a variable BsmtFinType2. *Numérico.*

BsmtUnfSF: Área (pies cuadrados) sin terminación en el sótano. *Numérico.*

TotalBsmtSF: Área (pies cuadrados) del sótano. *Numérico.*

Heating: Tipo de calefacción. *Polinomial.*

HeatingQC: Calidad y condición de la calefacción. *Polinomial.*

CentralAir: Aire acondicionado centralizado. *Binomial.*

Electrical: Sistema eléctrico. *Polinomial.*

1stFlrSF: Área del primer piso (pies cuadrados). *Numérico.*

2ndFlrSF: Área del segundo piso (pies cuadrados). *Numérico.*

LowQualFinSF: Área (pies cuadrados) de terminado de baja calidad (todos los pisos). *Numérico.*

GrLivArea: Área sobre el nivel del suelo, en pies cuadrados. *Numérico.*

BsmtFullBath: Baños completos en el sótano. *Binomial.*

BsmtHalfBath: Medio baño en el sótano. *Binomial.*

FullBath: Cantidad de baños completos sobre el nivel de suelo. *Numérico.*

HalfBath: Cantidad de medios baños sobre el nivel del suelo. *Numérico.*

Bedroom: Número de cuartos sobre el nivel del suelo. *Numérico.*

Kitchen: Número de cocinas sobre el nivel del suelo. *Numérico.*

KitchenQual: Calidad de la cocina. *Polinomial.*

TotRmsAbvGrd: Cantidad de habitaciones sobre el nivel del suelo. *Numérico.*

Functional: Calificación en funcionalidad del hogar. *Polinomial.*

Fireplaces: Cantidad de chimeneas. *Numérico.*

FireplaceQu: Calidad de la chimenea. *Polinomial.*

GarageType: Ubicación del garaje. *Polinomial.*

GarageYrBlt: Año de construcción del garaje. *Numérico.*

GarageFinish: Acabado interior del garaje. *Polinomial.*

GarageCars: Tamaño del garaje en la capacidad del automóviles. *Numérico.*

GarageArea: Tamaño del garaje en pies cuadrados. *Numérico.*

GarageQual: Calidad del garaje. *Polinomial.*

GarageCond: Condición del garaje. *Polinomial.*

PavedDrive: Entrada pavimentada. *Polinomial.*

WoodDeckSF: Área deck (de madera) en pies cuadrados. *Numérico.*

OpenPorchSF: Área de porche abierto en pies cuadrados. *Numérico.*

EnclosedPorch: Área de porche cerrado en pies cuadrados. *Numérico.*

3SsnPorch: Área de porche de tres estaciones en pies cuadrados. *Numérico.*

ScreenPorch: Área de porche de pantalla (cerrado con vidrios) en pies cuadrados. *Numérico.*

PoolArea: Área de la piscina, en pies cuadrados. *Numérico.*

PoolQC: Calidad de la piscina. *Polinomial.*

Fence: Calidad de la cerca. *Polinomial.*

MiscFeature: Característica miscelánea no cubierta en otras categorías (ej: ascensor). *Polinomial.*

MiscVal: Precio de la característica miscelánea (en dólares). *Numérico.*

MoSold: Mes de venta (MM). *Numérico.*

YrSold: Año de venta. *Numérico.*

SaleType: Tipo de venta. *Polinomial.*

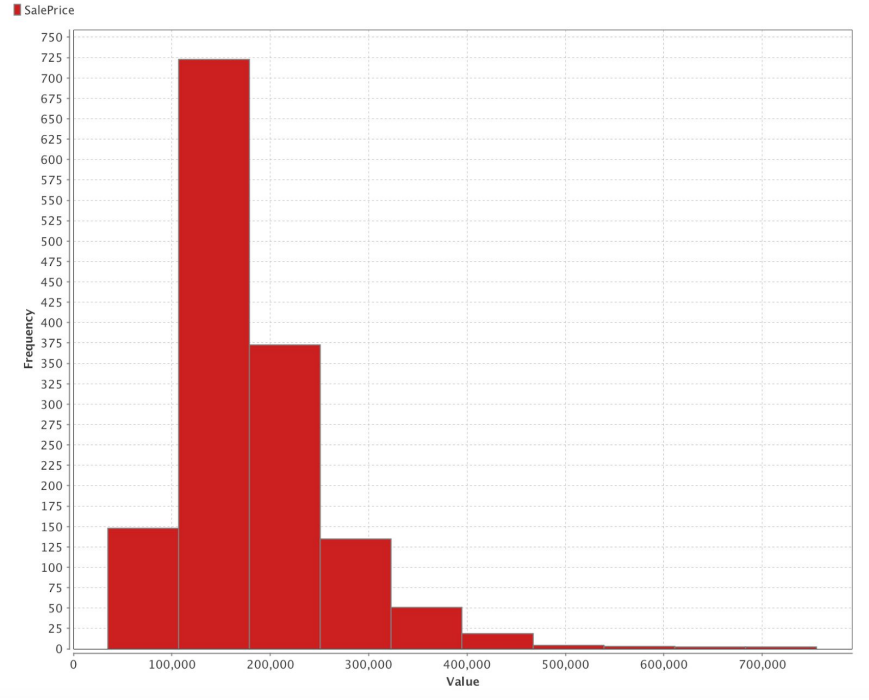
SaleCondition: Condición de venta. *Polinomial.*

Analizando la variable a predecir: SalePrice

Datos estadísticos

count	1460
mean	180921,19
std	79442,50
min	34900,00
25%	129975,00
50%	163000,00
75%	214000,00
max	755000,00

Histograma



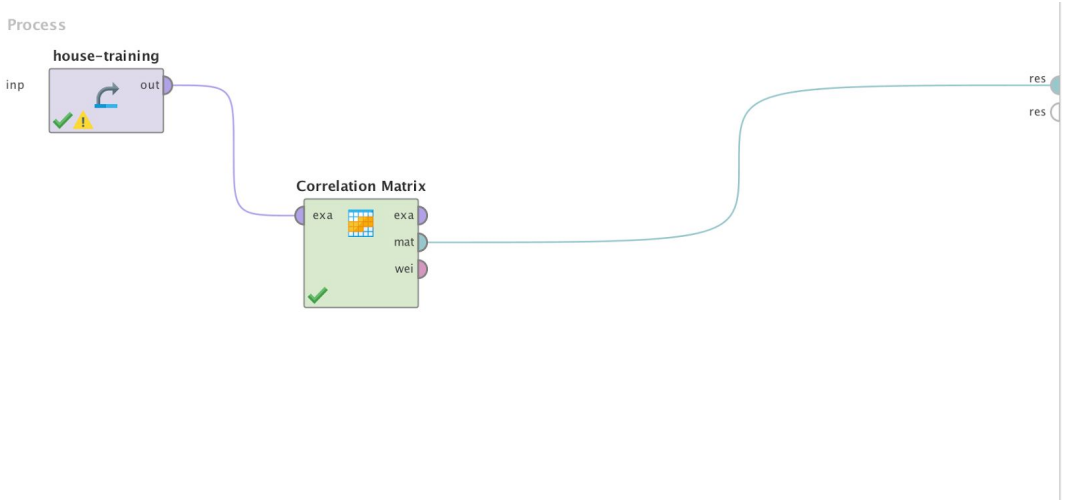
De la siguiente gráfica y resumen estadístico podemos observar principalmente que no contamos con registros donde nuestra variable a predecir tome el valor 0, se desvía de una distribución normal y tiene una asimetría positiva visible (1,882 una vez calculada).

En orden de poder llevar estos datos a un distribución normal se le aplicará la función logaritmo a la variable.

Relación entre variables dependientes e independiente

Al tratarse de un dataset con muchas variables dependientes, primero debemos considerar cuales pueden ser las variables más importante a la hora de intentar predecir nuestra variable de salida.

Para ello, utilizamos el step de rapidminer *Correlation Matrix,* generando así una matriz de correlación donde podamos observar cuales son las variables más fuertemente correlacionadas con *SalePrice*.



Attributes	SalePrice ↓	Id	MSSub...	MSZoni...	LotFro...	LotArea	Street	Alley
SalePrice	1	-0.022	-0.084	-0.116	0.203	0.264	-0.041	-0.093
OverallQual	0.791	-0.028	0.033	-0.015	0.203	0.106	-0.059	-0.027
GrLivArea	0.709	0.008	0.075	-0.061	0.152	0.263	-0.044	-0.004
GarageCars	0.640	0.017	-0.040	-0.077	0.147	0.155	0.020	-0.047
GarageArea	0.623	0.018	-0.099	-0.058	0.132	0.180	0.048	-0.063
TotalBsmtSF	0.614	-0.015	-0.239	-0.176	0.164	0.261	-0.005	-0.113
1stFlrSF	0.606	0.010	-0.252	-0.228	0.143	0.299	-0.006	-0.129
FullBath	0.561	0.006	0.132	0.011	0.149	0.126	-0.046	-0.005
TotRmsAbvGrd	0.534	0.027	0.040	-0.090	0.082	0.190	-0.047	-0.043
YearBuilt	0.523	-0.013	0.028	-0.063	0.169	0.014	-0.021	-0.172
YearRemodAdd	0.507	-0.022	0.041	0.033	0.176	0.014	-0.065	0.010
MasVnrArea	0.477	-0.050	0.023	-0.041	0.095	0.104	-0.017	-0.020
Fireplaces	0.467	-0.020	-0.046	-0.194	0.073	0.271	0.005	-0.114
FireplaceQu	0.403	-0.013	-0.050	-0.146	0.087	0.138	-0.026	-0.084
BsmtFinSF1	0.386	-0.005	-0.070	-0.150	0.065	0.214	0.016	-0.130
WoodDeckSF	0.324	-0.030	-0.013	-0.133	0.023	0.172	0.018	-0.117
2ndFlrSF	0.319	0.006	0.308	0.120	0.054	0.051	-0.047	0.104
OpenPorchSF	0.316	-0.000	-0.006	0.109	0.085	0.085	0.006	0.075

De acuerdo a la documentación del operador de RapidMiner, las correlaciones van desde -1 a 1 e indican cuán relacionadas están las variables. Si la correlación es cercana a 1 quiere decir la relación entre ambas es fuerte y que para valores altos de una variable vamos a tener valores altos de la otra y viceversa.

A partir de esta matriz podemos concluir que las variables que más correlación tienen con nuestra variable de salida (tomando una correlación de al menos 0,5) son:

1. OverallQual
2. GrLivArea
3. GarageCars
4. GarageArea
5. TotalBsmtSF
6. 1stFlrSF
7. FullBath
8. TotRmsAbvGrd
9. YearBuilt
10. YearRemodAdd

Con dichos atributos debemos comenzar a analizar la importancia de cada uno dentro de nuestro problema y si entre ellos están o no relacionados.

Mantenemos	Descartamos	Motivo
OverallQual		Altamente correlacionado e importante a la hora de predecir el precio de una casa
GrLivArea		Altamente correlacionado e importante a la hora de predecir el precio de una casa
GarageCars		Altamente correlacionado e importante a la hora de predecir el precio de una casa
	GarageArea	Altamente correlacionado e importante a la hora de predecir el precio de una casa, pero con una fuerte relación a la variable GarageCars (con una es suficiente). Mantenemos la de mayor correlación con la variable de salida
TotalBsmtSF		Altamente correlacionado e importante a la hora de predecir el precio de una casa
	1stFlrSF	Altamente correlacionado e importante a la hora de predecir el precio de una casa, pero con una fuerte relación a la variable TotalBsmtSF (con una es suficiente). Mantenemos la de mayor correlación con la variable de salida
FullBath		Altamente correlacionado e importante a la hora de predecir el precio de una casa
	TotRmsAbvGrd	Altamente correlacionado e importante a la hora de predecir el precio de una casa, pero con una fuerte relación a la variable GrLivArea. Mantenemos la de mayor correlación con la variable de salida
YearBuilt		Altamente correlacionado e importante a la hora de predecir el precio de una casa*
	YearRemodAdd	Altamente correlacionado e importante a la hora de predecir el precio de una casa, pero con una fuerte relación a la variable YearBuilt. Mantenemos la de mayor correlación con la variable de salida*

* Atributos de tipo fechas deben de tener consideraciones especiales a la hora de utilizar un modelo de predicción, como transformarlos a “antigüedad”. En este caso al ser únicamente el año el modelo no se ve afectado.

Missing data

Al tratar con atributos faltantes lo primero que debemos preguntarnos es qué tan importantes son estos atributos. A continuación detallamos las diferentes variables que tienen datos faltantes y como manejaremos cada uno, ya sea que al final lo utilicemos o no.

Name	Type	Missing	Statistics		Filter (19 / 81 attributes): Search for Attribute.
PoolQC	Polynominal	1453	Least NA (0)	Most Gd (3)	Values Gd (3), Ex (2), ...[2 more]
MiscFeature	Polynominal	1406	Least NA (0)	Most Shed (49)	Values Shed (49), Gar2 (2), ...[3 more]
Alley	Polynominal	1369	Least NA (0)	Most Grvl (50)	Values Grvl (50), Pave (41), ...[1 more]
Fence	Polynominal	1179	Least NA (0)	Most MnPrv (157)	Values MnPrv (157), GdPrv (59), ...[3 more]
FireplaceQu	Polynominal	690	Least NA (0)	Most Gd (380)	Values Gd (380), TA (313), ...[4 more]
LotFrontage	Polynominal	259	Least NA (0)	Most 60 (143)	Values 60 (143), 70 (70), ...[109 more]
GarageType	Polynominal	81	Least NA (0)	Most Attchd (870)	Values Attchd (870), Detchd (387), ...[5 more]
GarageYrBlt	Polynominal	81	Least NA (0)	Most 2005 (65)	Values 2005 (65), 2006 (59), ...[96 more]
GarageFinish	Polynominal	81	Least NA (0)	Most Unf (605)	Values Unf (605), RFn (422), ...[2 more]
GarageQual	Polynominal	81	Least NA (0)	Most TA (1311)	Values TA (1311), Fa (48), ...[4 more]
GarageCond	Polynominal	81	Least NA (0)	Most TA (1326)	Values TA (1326), Fa (35), ...[4 more]
BsmtExposure	Polynominal	38	Least NA (0)	Most No (953)	Values No (953), Av (221), ...[3 more]
BsmtFinType2	Polynominal	38	Least NA (0)	Most Unf (1256)	Values Unf (1256), Rec (54), ...[5 more]
BsmtQual	Polynominal	37	Least NA (0)	Most TA (649)	Values TA (649), Gd (618), ...[3 more]
BsmtCond	Polynominal	37	Least NA (0)	Most TA (1311)	Values TA (1311), Gd (65), ...[3 more]
BsmtFinType1	Polynominal	37	Least NA (0)	Most Unf (430)	Values Unf (430), GLQ (418), ...[5 more]
MasVnrType	Polynominal	8	Least NA (0)	Most None (864)	Values None (864), BrkFace (445), ...[3 more]
MasVnrArea	Polynominal	8	Least NA (0)	Most 0 (861)	Values 0 (861), 108 (8), ...[326 more]
Electrical	Polynominal	1	Least NA (0)	Most SBrkr (1334)	Values SBrkr (1334), FuseA (94), ...[4 more]

En primer caso, decidimos remover las columnas que presentan más de un 17% de atributos faltantes en el dataset (recordemos que el mismo tiene 1406 observaciones), esto incluye *PoolQC*, *MiscFeature*, *Alley*, *Fence*, *FireplaceQu* y *LotFrontage*. Además no consideremos sean variables importantes a la hora de predecir el precio de una casa.

El resto de los atributos presentan un porcentaje de atributos faltantes bajo.

En el caso de las variables relacionadas a Garage, presentan la misma cantidad de atributos faltantes, y viendo en mayor detalle estos registros podemos observar que estos datos faltantes pertenecen a las mismas filas del dataset. Al tratarse de 81 filas (5%) decidimos remover las mismas del dataset.

Lo mismo ocurre con las variables relacionadas a *Bsmt* y aplicamos el mismo tratamiento.

MasVnrArea y *MasVnrType* presentan una fuerte correlación con *YearBuilt* y *OverallQual,* ambas ya incluidas en nuestros atributos a considerar, además de tampoco considerarlas de gran importancia a la hora de predecir nuestra variable de salida, por lo que podemos descartar utilizar ambas variables.

Finalmente tenemos un registro con datos faltantantes en *Electrical*. Al tratarse de una única fila podemos eliminar dicha observación.

Outliers

Para la detección de outliers vamos a usar el operador de RapidMiner llamado Detect Outliers por distancia ya que son todos atributos numéricos. Este atributo si bien simplifica la tarea de detección de outliers tiene una pequeña desventaja que es que le tenemos que decir la cantidad de outliers que queremos que nos detecte.

Aplicando PCA para reducir la dimensionalidad de la data, podemos comenzar a ver en una primera instancia cuantos posibles registros outliers tenemos en nuestro dataset. Una vez aplicado PCA, podemos ver a través de una gráfica de dispersión de puntos que aproximadamente y a grandes rasgos, existen 5 ejemplos que están fuera de los parámetros normales.