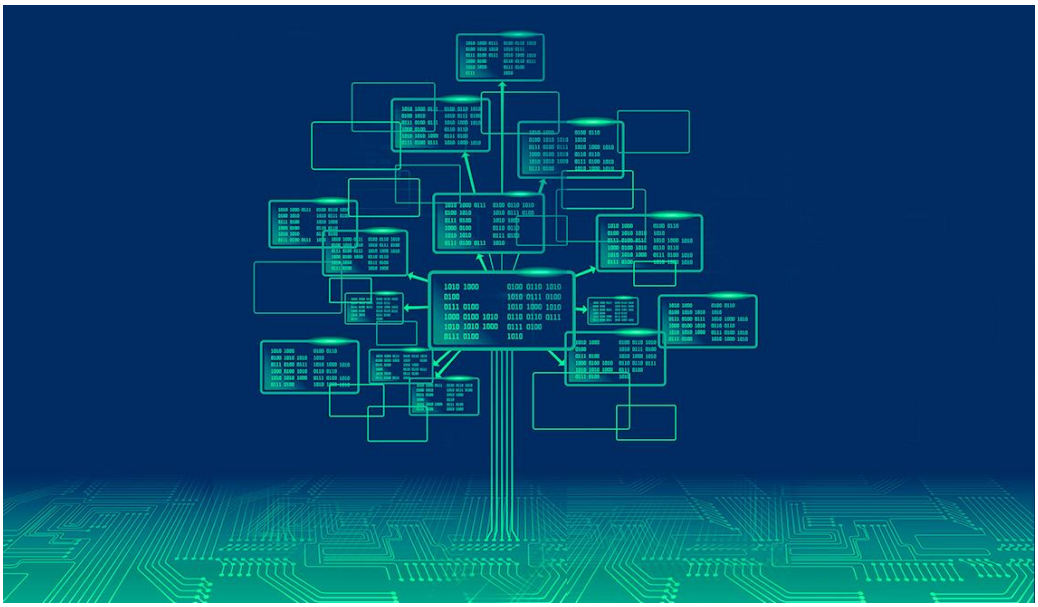
Universidad del Valle de Guatemala

Minería de Datos

HOJA DE TRABAJO #3
ARBOLES DE DECISIÓN



16/03/2020

Análisis de decisión sobre elección de vivienda

En el siguiente documento se presentan los diferentes algoritmos que caracterizan el "árbol de decisión". Muchas veces ignoramos factores esenciales al momento de comprar una casa. En la presente investigación se mostrará cómo es que se debe de trabajar y que factores analizar al momento de hacer la compra, el alquiler o la remodelación de una casa.

Pablo Sao – 11530

Mercedes Retolaza - 16339

Contenido

INTRODUCCIÓN

Aprendizaje basado en árboles de decisión utiliza un árbol de decisión como un modelo predictivo que mapea observaciones sobre un artículo a conclusiones sobre el valor objetivo del artículo. Es uno de los enfoques de modelado predictivo utilizadas en estadísticas, minería de datos y aprendizaje automático. Los modelos de árbol, donde la variable de destino puede tomar un conjunto finito de valores se denominan árboles de clasificación. En estas estructuras de árbol, las hojas representan etiquetas de clase y las ramas representan las conjunciones de características que conducen a esas etiquetas de clase. Los árboles de decisión, donde la variable de destino puede tomar valores continuos (por lo general números reales) se llaman árboles de regresión.

En análisis de decisión, un árbol de decisión se puede utilizar para representar visualmente y de forma explícita decisiones y toma de decisiones. En minería de datos, un árbol de decisión describe datos, pero no las decisiones; más bien el árbol de clasificación resultante puede ser un usado como entrada para la toma de decisiones. Esta página se ocupa de los árboles de decisión en la minería de datos.

ANÁLISIS

En el siguiente espacio se definirán el proceso que se tuvo que llevar a cabo para poder obtener los resultados al momento de preparar los datos.

El siguiente documento encontrará las gráficas relevantes, descripciones e interpretación de los datos obtenidos. Se adjuntará un documento en Python (JupyterNotebook) donde podrá visualizar los algoritmos y procedimiento que se utilizó para llevar a cabo los análisis que nos llevaron a nuestros resultados.

Variables

Las variables que representan un valor significativo son todas aquellas que nos ayudarán a obtener los valores esperados en los gráficos. Pero también existen varios datos que nunca vamos a utilizar porque en este momento no estamos realizando un análisis que los involucre.

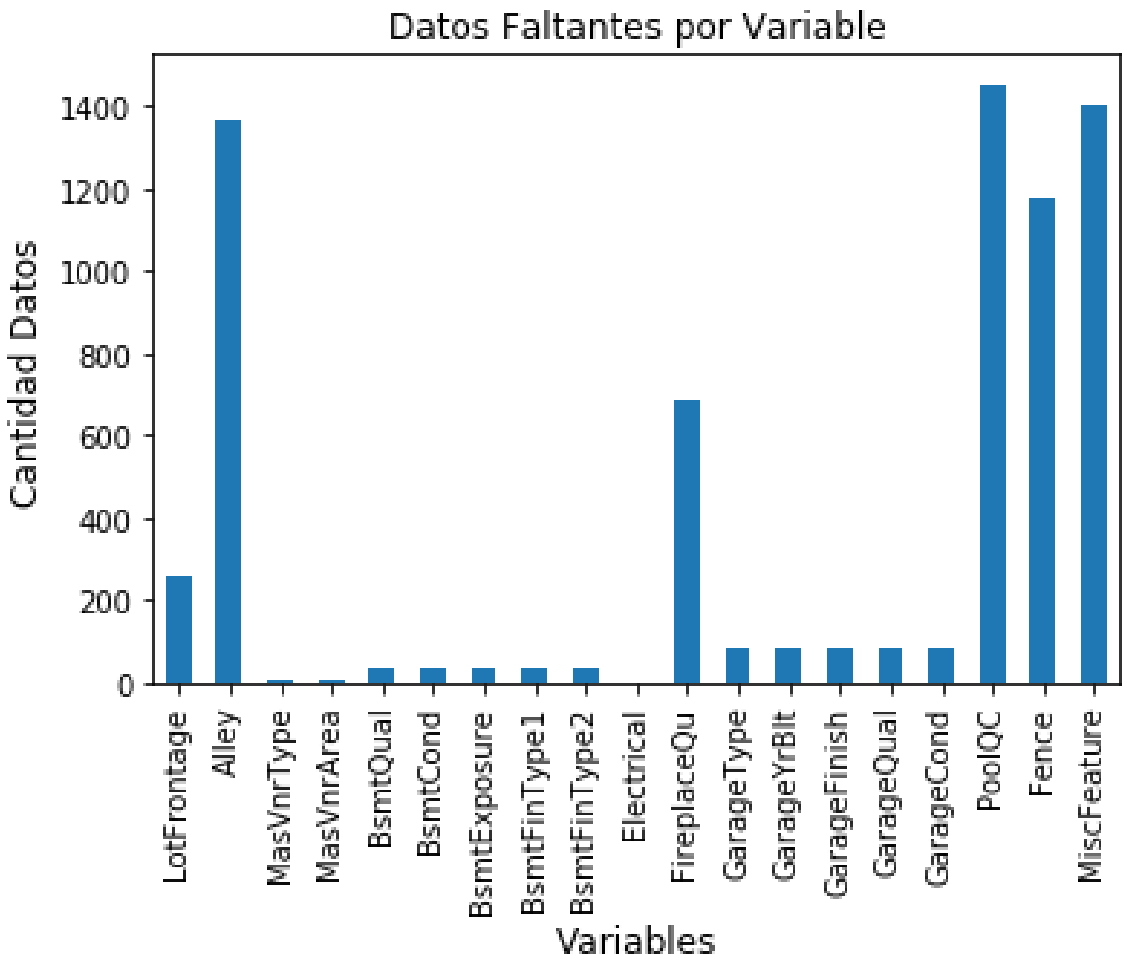
Con los datos provistos, se adjunta una descripción de las variables que se tenían con sus posibles valores, en la cual se basó para la transformación y el análisis de los datos. Dicha descripción se encuentra junto con el repositorio y en la sección de anexos del presente documento.

RESUTADOS

Se realizó una limpieza de datos general por la base de datos en donde se extrajo la información. El proceso que se realizó fue un análisis exploratorio por cada uno de los datos. De esos se tomó únicamente los que íbamos a analizar. Aparte se realizó una actualización de variables de acorde a los valores numéricos que se utilizarán.

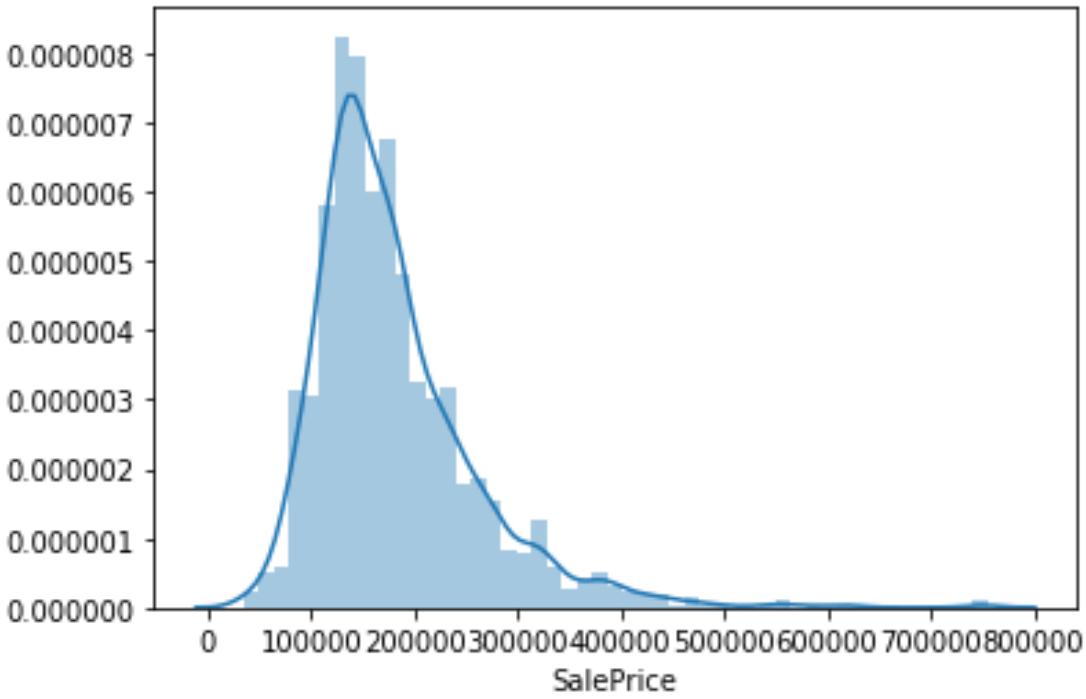
También se evalúa cuantos datos hacen falta por variable. Para resolverlo se llenan los datos con la moda estadística de los valores, se cambia codificación, por ejemplo, si tiene cerco originalmente tenía NaN y se cambia a None, para que sea más entendible.

Datos faltantes por variable



Descripción: En la siguiente gráfica se tienen los datos de las variables de la fuente de información que no poseen datos. En base a esta gráfica podemos tener una idea de que variables deben ser trabajadas para completar la información. Tomando un caso en específico, como el si una casa tiene piscina, llenaremos los datos faltantes con la opción que nos provee la fuente de información, colocando que la vivienda no tiene dicha característica, en otros casos tomaremos el valor que se repite con mayor frecuencia (moda estadística) para completar la información, de tal forma que está grafica no muestre datos.

Verificación de datos



Descripción: Luego de completar la información, procedemos a realizar un análisis del precio, ya que este será el valor que tomaremos como eje Y para realizar el cruce con las características de la vivienda, en el momento de analizar el algoritmo. Logrando aprecias que es leptocúrtica, concentrándose el precio, concentrándose en el rango de $100,000.00 y $300,000.00, por lo que los datos no tendrán una dispersión alta.

- **Distribución:** La definición de distribución se relaciona al conjunto de acciones que se llevan a cabo desde que un producto se elabora por parte del fabricante hasta que es comprado por el consumidor final. El objetivo de la distribución es garantizar la llegada de un producto o bien hasta el cliente.

  El concepto de distribución resulta muy importante para garantizar las ventas de un producto, ya que no resulta suficiente con tener un artículo de calidad y a un precio competitivo. También es preciso que sea accesible para los usuarios. De ahí que haya que colocar el producto en los puntos de venta habilitados.

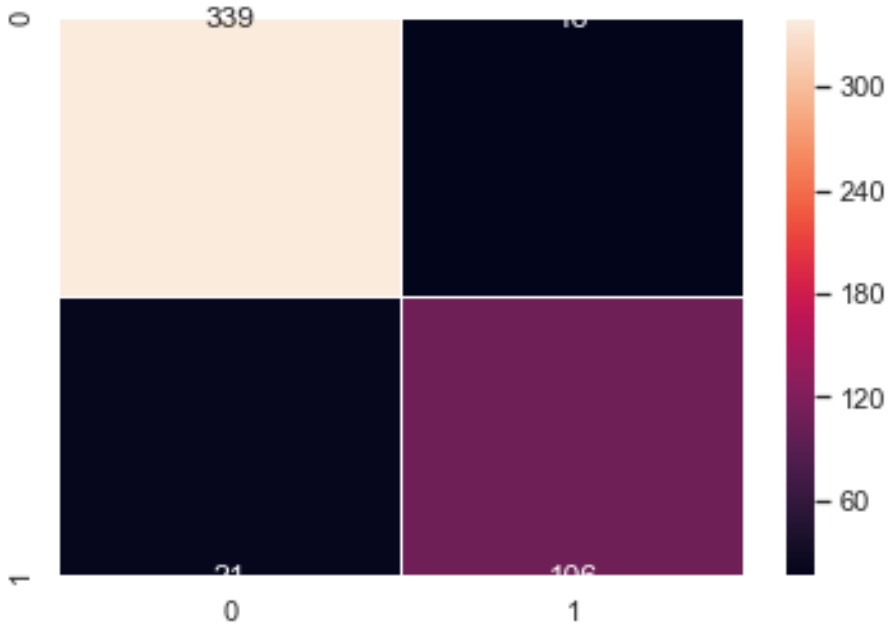
Análisis de cruces de variables

El cruce de variables pretende identificar si existe relación entre dos o más de ellas, además, de posibilitar el análisis de estas variables en una sola tabla, en lugar de construir dos cuadros simples. Estas tablas aplican fundamentalmente para variables categóricas o incluso cuantitativa discreta si ésta no tiene muchas categorías de respuestas.

Descripción: Se despliega los diferentes resultados para un cruce de variables, esto nos orienta a las variables que debemos de utilizar para obtener un análisis correcto de datos y además obtener/verificar que nuestras variables de uso sean correctas. Es decir, que sean interpretables por nuestros algoritmos a utilizar.

Random Forest

Es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. Es una modificación sustancial de bagging que construye una larga colección de árboles no correlacionados y luego los promedia.

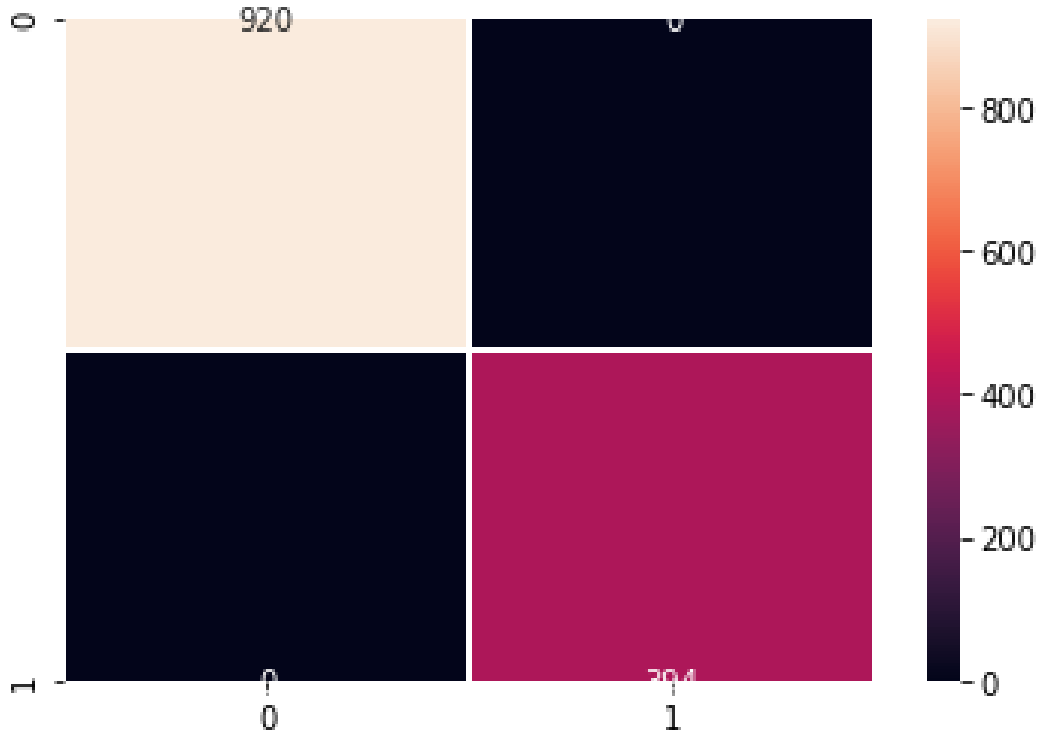


Los valores que se encuentran en la gráfica son los siguientes:

En la siguiente tabla se muestra los valores de la matriz de confusión y los datos de clasificación que estamos trabajando en dicho gráfico.

```
=== Confusion Matrix ===
[[339  16]
 [ 21 106]]

=== Classification Report ===
              precision    recall  f1-score   support

           0       0.94      0.95      0.95       355
           1       0.87      0.83      0.85       127

    accuracy                           0.92       482
   macro avg       0.91      0.89      0.90       482
weighted avg       0.92      0.92      0.92       482


=== All AUC Scores ===
[0.98848499 0.97538948 0.98724317 0.98351772 0.97538948 0.99277489
 0.97155114 0.97115385 0.98534799 0.97149725]

=== Mean AUC Score ===
Mean AUC Score - Random Forest:  0.9802349953519801
```

Matriz de confusión



Estos son los valores de la matriz:

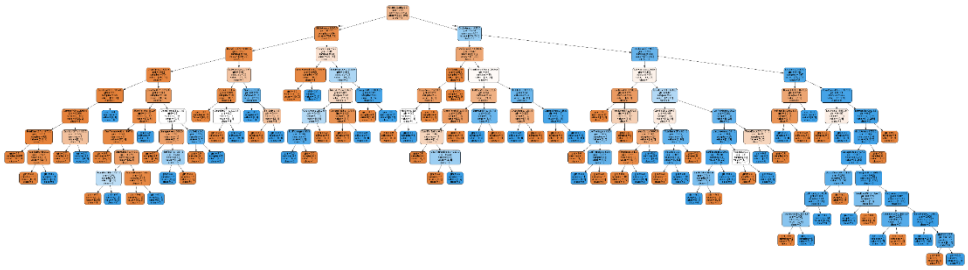
```
[[920   0] [ 0 394]]
```

Descripción de resultados de matriz de confusión:

- Puntaje precisión: 1.0 --> tp/tp+fp
  -
- Puntaje Exactitud: 1.0--> total correct
  -
- Puntaje de Llamada: 1.0--> tp/tp+fn
  -
- Puntaje Datos Test score: 1.0
  -

Descripción: En estos resultados podemos identificar que la matriz de confusión es más preciosa a la del árbol de decisión. Esto se debe a que al delimitar los resultados hacemos un análisis más detallado y así determinamos que tan confiable es el algoritmo para ser utilizado en la predicción.

Árbol de decisiones armado



Se adjuntará imagen de árbol de decisiones, en la cual se basa para tomar las decisiones, y dar un valor en base a las características de las viviendas.

ANEXOS

Información de las variables y tipo de datos de la fuente de información del precio de viviendas.

MSSubClass: Identifies the type of dwelling involved in the sale.

20	1-STORY 1946 & NEWER ALL STYLES
30	1-STORY 1945 & OLDER
40	1-STORY W/FINISHED ATTIC ALL AGES
45	1-1/2 STORY - UNFINISHED ALL AGES
50	1-1/2 STORY FINISHED ALL AGES
60	2-STORY 1946 & NEWER
70	2-STORY 1945 & OLDER
75	2-1/2 STORY ALL AGES
80	SPLIT OR MULTI-LEVEL
85	SPLIT FOYER
90	DUPLEX - ALL STYLES AND AGES
120	1-STORY PUD (Planned Unit Development) - 1946 & NEWER
150	1-1/2 STORY PUD - ALL AGES
160	2-STORY PUD - 1946 & NEWER
180	PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
190	2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale. -- X

A	Agriculture
C	Commercial
FV	Floating Village Residential
I	Industrial
RH	Residential High Density
RL	Residential Low Density

RP Residential Low Density Park
RM Residential Medium Density

LotFrontage: Linear feet of street connected to property -- X

LotArea: Lot size in square feet

Street: Type of road access to property -- X

Grvl Gravel
Pave Paved

Alley: Type of alley access to property -- X

Grvl Gravel
Pave Paved
NA No alley access

LotShape: General shape of property

Reg Regular
IR1 Slightly irregular
IR2 Moderately Irregular
IR3 Irregular

LandContour: Flatness of the property

Lvl Near Flat/Level
Bnk Banked - Quick and significant rise from street grade to building
HLS Hillside - Significant slope from side to side
Low Depression

Utilities: Type of utilities available --- X

AllPub All public Utilities (E,G,W,& S)
NoSewr Electricity, Gas, and Water (Septic Tank)
NoSeWa Electricity and Gas Only
ELO Electricity only

LotConfig: Lot configuration --- X

Inside Inside lot
Corner Corner lot
CulDSac Cul-de-sac
FR2 Frontage on 2 sides of property
FR3 Frontage on 3 sides of property

LandSlope: Slope of property

Gtl Gentle slope

Mod	Moderate Slope
Sev	Severe Slope

Neighborhood: Physical locations within Ames city limits

Blmngtn	Bloomington Heights
Blueste	Bluestem
BrDale	Briardale
BrkSide	Brookside
ClearCr	Clear Creek
CollgCr	College Creek
Crawfor	Crawford
Edwards	Edwards
Gilbert	Gilbert
IDOTRR	Iowa DOT and Rail Road
MeadowV	Meadow Village
Mitchel	Mitchell
Names	North Ames
NoRidge	Northridge
NPkVill	Northpark Villa
NridgHt	Northridge Heights
NWAmes	Northwest Ames
OldTown	Old Town
SWISU	South & West of Iowa State University
Sawyer	Sawyer
SawyerW	Sawyer West
Somerst	Somerset
StoneBr	Stone Brook
Timber	Timberland
Veenker	Veenker

Condition1: Proximity to various conditions

Artery	Adjacent to arterial street
Feedr	Adjacent to feeder street
Norm	Normal
RRNn	Within 200' of North-South Railroad
RRAn	Adjacent to North-South Railroad
PosN	Near positive off-site feature--park, greenbelt, etc.
PosA	Adjacent to postive off-site feature
RRNe	Within 200' of East-West Railroad
RRAe	Adjacent to East-West Railroad

Condition2: Proximity to various conditions (if more than one is present)

Artery	Adjacent to arterial street
Feedr	Adjacent to feeder street
Norm	Normal
RRNn	Within 200' of North-South Railroad
RRAn	Adjacent to North-South Railroad

PosN	Near positive off-site feature--park, greenbelt, etc.
PosA	Adjacent to postive off-site feature
RRNe	Within 200' of East-West Railroad
RRAe	Adjacent to East-West Railroad

BldgType: Type of dwelling

1Fam	Single-family Detached
2FmCon	Two-family Conversion; originally built as one-family dwelling
Duplx	Duplex
TwnhsE	Townhouse End Unit
TwnhsI	Townhouse Inside Unit

HouseStyle: Style of dwelling --

1Story	One story
1.5Fin	One and one-half story: 2nd level finished
1.5Unf	One and one-half story: 2nd level unfinished
2Story	Two story
2.5Fin	Two and one-half story: 2nd level finished
2.5Unf	Two and one-half story: 2nd level unfinished
SFoyer	Split Foyer
SLvl	Split Level

OverallQual: Rates the overall material and finish of the house

10	Very Excellent
9	Excellent
8	Very Good
7	Good
6	Above Average
5	Average
4	Below Average
3	Fair
2	Poor
1	Very Poor

OverallCond: Rates the overall condition of the house

10	Very Excellent
9	Excellent
8	Very Good
7	Good
6	Above Average
5	Average
4	Below Average
3	Fair
2	Poor
1	Very Poor

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

RoofStyle: Type of roof -- X

Flat	Flat
Gable	Gable
Gambrel	Gabrel (Barn)
Hip	Hip
Mansard	Mansard
Shed	Shed

RoofMatl: Roof material -- X

ClyTile	Clay or Tile
CompShg	Standard (Composite) Shingle
Membran	Membrane
Metal	Metal
Roll	Roll
Tar&Grv	Gravel & Tar
WdShake	Wood Shakes
WdShngl	Wood Shingles

Exterior1st: Exterior covering on house -- X

AsbShng	Asbestos Shingles
AsphShn	Asphalt Shingles
BrkComm	Brick Common
BrkFace	Brick Face
CBlock	Cinder Block
CemntBd	Cement Board
HdBoard	Hard Board
ImStucc	Imitation Stucco
MetalSd	Metal Siding
Other	Other
Plywood	Plywood
PreCast	PreCast
Stone	Stone
Stucco	Stucco
VinylSd	Vinyl Siding
Wd Sdng	Wood Siding
WdShing	Wood Shingles

Exterior2nd: Exterior covering on house (if more than one material)

AsbShng	Asbestos Shingles
AsphShn	Asphalt Shingles
BrkComm	Brick Common
BrkFace	Brick Face

CBlock Cinder Block
CemntBd Cement Board
HdBoard Hard Board
ImStucc Imitation Stucco
MetalSd Metal Siding
Other Other
Plywood Plywood
PreCast PreCast
Stone Stone
Stucco Stucco
VinylSd Vinyl Siding
Wd Sdng Wood Siding
WdShing Wood Shingles

MasVnrType: Masonry veneer type

BrkCmn Brick Common
BrkFace Brick Face
CBlock Cinder Block
None None
Stone Stone

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

Ex Excellent
Gd Good
TA Average/Typical
Fa Fair
Po Poor

ExterCond: Evaluates the present condition of the material on the exterior

Ex Excellent
Gd Good
TA Average/Typical
Fa Fair
Po Poor

Foundation: Type of foundation -- X

BrkTil Brick & Tile
CBlock Cinder Block
PConc Poured Contrete
Slab Slab
Stone Stone
Wood Wood

BsmtQual: Evaluates the height of the basement

Ex Excellent (100+ inches)
Gd Good (90-99 inches)
TA Typical (80-89 inches)
Fa Fair (70-79 inches)
Po Poor (<70 inches
NA No Basement

BsmtCond: Evaluates the general condition of the basement

Ex Excellent
Gd Good
TA Typical - slight dampness allowed
Fa Fair - dampness or some cracking or settling
Po Poor - Severe cracking, settling, or wetness
NA No Basement

BsmtExposure: Refers to walkout or garden level walls

Gd Good Exposure
Av Average Exposure (split levels or foyers typically score average or above)
Mn Mimimum Exposure
No No Exposure
NA No Basement

BsmtFinType1: Rating of basement finished area

GLQ Good Living Quarters
ALQ Average Living Quarters
BLQ Below Average Living Quarters
Rec Average Rec Room
LwQ Low Quality
Unf Unfinshed
NA No Basement

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types)

GLQ Good Living Quarters
ALQ Average Living Quarters
BLQ Below Average Living Quarters
Rec Average Rec Room
LwQ Low Quality
Unf Unfinshed
NA No Basement

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

Floor	Floor Furnace
GasA	Gas forced warm air furnace
GasW	Gas hot water or steam heat
Grav	Gravity furnace
OthW	Hot water or steam heat other than gas
Wall	Wall furnace

HeatingQC: Heating quality and condition

Ex	Excellent
Gd	Good
TA	Average/Typical
Fa	Fair
Po	Poor

CentralAir: Central air conditioning

N	No
Y	Yes

Electrical: Electrical system -- X

SBrkr	Standard Circuit Breakers & Romex
FuseA	Fuse Box over 60 AMP and all Romex wiring (Average)
FuseF	60 AMP Fuse Box and mostly Romex wiring (Fair)
FuseP	60 AMP Fuse Box and mostly knob & tube wiring (poor)
Mix	Mixed

1stFlrSF: First Floor square feet -- X

2ndFlrSF: Second floor square feet -- X

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: Bedrooms above grade (does NOT include basement bedrooms)

Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

Ex Excellent
Gd Good
TA Typical/Average
Fa Fair
Po Poor

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

Functional: Home functionality (Assume typical unless deductions are warranted)

Typ Typical Functionality
Min1 Minor Deductions 1
Min2 Minor Deductions 2
Mod Moderate Deductions
Maj1 Major Deductions 1
Maj2 Major Deductions 2
Sev Severely Damaged
Sal Salvage only

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

Ex Excellent - Exceptional Masonry Fireplace
Gd Good - Masonry Fireplace in main level
TA Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement
Fa Fair - Prefabricated Fireplace in basement
Po Poor - Ben Franklin Stove
NA No Fireplace

GarageType: Garage location

2Types More than one type of garage
Attchd Attached to home
Basment Basement Garage
BuiltIn Built-In (Garage part of house - typically has room above garage)
CarPort Car Port
Detchd Detached from home
NA No Garage

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

Fin Finished
RFn Rough Finished
Unf Unfinished
NA No Garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

Ex Excellent
Gd Good
TA Typical/Average
Fa Fair
Po Poor
NA No Garage

GarageCond: Garage condition

Ex Excellent
Gd Good
TA Typical/Average
Fa Fair
Po Poor
NA No Garage

PavedDrive: Paved driveway

Y Paved
P Partial Pavement
N Dirt/Gravel

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

Ex Excellent
Gd Good
TA Average/Typical

Fa Fair
NA No Pool

Fence: Fence quality

GdPrv Good Privacy
MnPrv Minimum Privacy
GdWo Good Wood
MnWw Minimum Wood/Wire
NA No Fence

MiscFeature: Miscellaneous feature not covered in other categories

Elev Elevator
Gar2 2nd Garage (if not described in garage section)
Othr Other
Shed Shed (over 100 SF)
TenC Tennis Court
NA None

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: Type of sale

WD Warranty Deed - Conventional
CWD Warranty Deed - Cash
VWD Warranty Deed - VA Loan
New Home just constructed and sold
COD Court Officer Deed/Estate
Con Contract 15% Down payment regular terms
ConLw Contract Low Down payment and low interest
ConLI Contract Low Interest
ConLD Contract Low Down
Oth Other

SaleCondition: Condition of sale

Normal Normal Sale
Abnorml Abnormal Sale - trade, foreclosure, short sale
AdjLand Adjoining Land Purchase
Alloca Allocation - two linked properties with separate deeds, typically condo with a garage unit
Family Sale between family members
Partial Home was not completed when last assessed (associated with New Homes)

Alquiler:
- Ubicación
- Requerimiento Basico:
  - cocina
  - lavanderia
  - parqueos
  - 3 cuartos
  - sala
  - comedor

Compra:
- Ubicacion
- Sala Familiar
- Estudio