HarvardX PH125.9xData Science: Capstone Choose Your Own!(House Pricing)

Jose Quesada

27/12/2020

Introduccion

This project consists of determining the value of a house according to its location, property characteristics and payment methods using the data set from kaggle House Prices - Advanced Regression Techniques https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data (kaggle competitions download -c house-prices-advanced-regression-techniques).

Importing Data & EDA

Files:

```
train.csv - the training set
test.csv - the test set
data_description.txt - full description of each column, originally prepared by Dean De Cock but lightly
sample_submission.csv - a benchmark submission from a linear regression on year and month of sale, lot s
```

```
read_csv <- function(file){
    path_data <- "data"
    filename <- paste(path_data,file,sep="/")
    csv__ <- read.csv(filename)
    csv__
}

test_set <-read_csv('test.csv')
train_set<- read_csv('train.csv')


#Train SET INFO
colnames_trian_set<-colnames(train_set)
memory_usage_train_set<-format(object.size(train_set),units="MB")
dim_train_set<- dim(train_set)

#TEST SET INFO
colnames_test_set<-colnames(test_set)
memory_usage_test_set<-format(object.size(test_set),units="MB")
dim_test_set <- dim(test_set)
```

train_set:

```
* Dimensions: 1460, 81
* Memory Usage: 0.7 Mb
```

test_set:

```
* Dimensions: 1459, 80
* Memory Usage: 0.7 Mb
```

For this project we going to join train and set data for the cleansing and EDA, later we going to split again by SalesPrices not null as train set and test set is null.

Comparing amount of columns between each dataset we can see that we have 1 more column in the train set vs the test set. **SalePrice** is the additional column in the train set and our **target value** for this model We going to use the train set to predict **SalePrice** on the test, first we going to make some EDA and data cleaning.

Total categorical columns: 43

Categorical Columns				
MSZoning	Street	Alley	LotShape	LandContour
Utilities	LotConfig	LandSlope	Neighborhood	Condition1
Condition2	BldgType	HouseStyle	RoofStyle	RoofMatl
Exterior1st	Exterior2nd	MasVnrType	ExterQual	ExterCond
Foundation	BsmtQual	BsmtCond	BsmtExposure	BsmtFinType1
BsmtFinType2	Heating	HeatingQC	CentralAir	Electrical
KitchenQual	Functional	FireplaceQu	GarageType	GarageFinish
GarageQual	GarageCond	PavedDrive	PoolQC	Fence
MiscFeature	SaleType	SaleCondition	MSZoning	Street
Alley	LotShape	LandContour	Utilities	LotConfig

Total numeric columns: 38

Numerical Columns			
Id	MSSubClass	LotFrontage	LotArea
OverallQual	OverallCond	YearBuilt	YearRemodAdd
MasVnrArea	BsmtFinSF1	BsmtFinSF2	BsmtUnfSF
TotalBsmtSF	X1stFlrSF	X2ndFlrSF	LowQualFinSF
GrLivArea	BsmtFullBath	BsmtHalfBath	FullBath
HalfBath	BedroomAbvGr	KitchenAbvGr	TotRmsAbvGrd
Fireplaces	GarageYrBlt	GarageCars	GarageArea
WoodDeckSF	OpenPorchSF	EnclosedPorch	X3SsnPorch
ScreenPorch	PoolArea	MiscVal	MoSold
YrSold	SalePrice	Id	MSSubClass

Im going to handle MSSubClass as categorical data although it is shown as numeric column, is actually a categorical data, the numbers is the columns are the type of dwelling involved in the sale, im removing from numeric columns and append as categorical.

MSSubClass:

```
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
```

Missing Values

For this analysis we going to select just the columns that have missing values, **if they not in plot or table its because they not have missing values**.

Description of columns with Missing Values:

Missing Categorical Columns	
name	prc_na
PoolQC	0.9965742
MiscFeature	0.9640288
Alley	0.9321686
Fence	0.8043851
FireplaceQu	0.4864680
GarageFinish	0.0544707
GarageQual	0.0544707
GarageCond	0.0544707
GarageType	0.0537855
BsmtCond	0.0280918
BsmtExposure	0.0280918
BsmtQual	0.0277492
BsmtFinType2	0.0274066
BsmtFinType1	0.0270641
MasVnrType	0.0082220
MSZoning	0.0013703
Utilities	0.0006852
Functional	0.0006852
Exterior1st	0.0003426
Exterior2nd	0.0003426
Electrical	0.0003426
KitchenQual	0.0003426
SaleType	0.0003426

```
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

Alley: Type of alley access to property

       Grvl Gravel
       Pave Paved
       NA   No alley access

Fence: Fence quality

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

GarageFinish: Interior finish of the garage

       Fin  Finished
       RFn  Rough Finished
       Unf  Unfinished
       NA   No Garage

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

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

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

MasVnrType: Masonry veneer type

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

Electrical: Electrical system

   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
```

Missing Numerical Columns	
name	prc_na
SalePrice	0.4998287
LotFrontage	0.1664954
GarageYrBlt	0.0544707
MasVnrArea	0.0078794
BsmtFullBath	0.0006852
BsmtHalfBath	0.0006852
BsmtFinSF1	0.0003426
BsmtFinSF2	0.0003426
BsmtUnfSF	0.0003426
TotalBsmtSF	0.0003426
GarageCars	0.0003426
GarageArea	0.0003426

```
LotFrontage: Linear feet of street connected to property
GarageYrBlt: Year garage was built
MasVnrArea: Masonry veneer area in square feet
```

```
evaluation_quality <- data.table("quality" = c("NA","Po","Fa","TA","Gd","Ex"), "score"=c(-999999,0,1,2,3
quality_columns<-names(train_set[,(grepl("Qu|Qua|QC|Cond",names(train_set)))&names(train_set) %in% categ
for(col in quality_columns){
  train_set[,col]<- mapvalues(as.vector(train_set[,col]),evaluation_quality$quality,evaluation_quality$s
  train_set[,col][is.na(train_set[col])]<--9999999
}
```

```
## The following `from` values were not present in `x`: NA, Po, Fa, TA, Gd, Ex
```

```
## The following `from` values were not present in `x`: NA, Po, Fa, TA, Gd, Ex
## The following `from` values were not present in `x`: NA, Po
## The following `from` values were not present in `x`: NA
## The following `from` values were not present in `x`: NA, Po
## The following `from` values were not present in `x`: NA, Ex
## The following `from` values were not present in `x`: NA
## The following `from` values were not present in `x`: NA, Po
## The following `from` values were not present in `x`: NA
## The following `from` values were not present in `x`: NA
## The following `from` values were not present in `x`: NA
## The following `from` values were not present in `x`: NA, Po, TA
## The following `from` values were not present in `x`: NA, Po, Fa, TA, Gd, Ex
```

ANEXO

```
## DATASET DETAILS
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.

       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

LotArea: Lot size in square feet
```

```
Street: Type of road access to property

       Grvl Gravel
       Pave Paved

Alley: Type of alley access to property

       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

       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

       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

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

RoofMatl: Roof material

       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

       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

```
       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

       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

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: 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 u
Family   Sale between family members
Partial  Home was not completed when last assessed (associated with New Homes)
```