PROPERTY PRICE PREDICTION USING LINEAR REGRESSION

By Darshan D S

Supervisor, Dr. Vinod Kumar Murti

1. Introduction

1.1. Abstract

A key challenge for property sellers is to determine the sale price of the property. The ability to predict the exact property value is beneficial for property investors as well as for buyers to plan their finances according to the price trend. The property prices depend on the number of features like the property area, basement square footage, year built, number of bedrooms, and so in.

1.2. Data

The data used for this project was compiled by Dean De Cock for use in data science education. There are 80 variables which focus on the quality and quantity of many physical attributes of the property. Most of the variables are exactly the type of information that a typical home buyer would want to know about a potential property (e.g., When was it built? How big is the lot? How many square feet of living space is in the dwelling? Is the basement finished? How many bathrooms are there?).

In general, the 20 continuous variables relate to various area dimensions for each observation. In addition to the typical lot size and total dwelling square footage found on most common home listings, other more specific variables are quantified in the dataset. Area measurements on the basement, main living area, and even porches are broken down into individual categories based on quality and type.

The 14 discrete variables typically quantify the number of items occurring within the house. Most are specifically focused on the number of kitchens, bedrooms, and bathrooms (full and half) located in the basement and above grade (ground) living area of the home. Additionally, the garage capacity and construction/remodeling dates are also recoded.

There are a large number of categorical variables (23 nominal, 23 ordinal) associated with this data set. They range from 2 to 28 classes with the smallest being STREET (gravel or paved) and the largest being NEIGHBORHOOD (areas within the city limits). The nominal variables typically identify various types of

dwellings, garages, materials, and environmental conditions while the ordinal variables typically rate various items within the property.

There are 2073 entries in the data set, where some of the variables have null values in it.

2. Software

The program used for this project is python. Python is a general-purpose programming language with lots of libraries which is very useful for data analysis and machine learning projects. The python version used for this project is Python 3.8.3.

The major libraries used for this project are mentioned below.

Libraries	Description	Developer
NumPy	NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.	Travis Oliphant
Pandas	Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.	Wes McKinney
Matplotlib	Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.	John D. Hunter
Seaborn	Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.	Michael Waskom
SciPy	SciPy is a free and open-source Python library used for scientific computing and technical computing. SciPy contains modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers and other tasks common in science and engineering.	Enthought
Scikit-learn	Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, ...	David Cournapeau

The Integrated Development Environment (IDE) used for this project is Jupyter notebook. The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

3. Data Description

The dataset contains 80 independent variables and a response variable which is property sale price. The detailed description of all the variables is given below.

1. Dwell_Type: Identifies the type of dwelling involved in the sale
    a. 20 1-STORY 1946 & NEWER ALL STYLES
    b. 30 1-STORY 1945 & OLDER
    c. 40 1-STORY W/FINISHED ATTIC ALL AGES
    d. 45 1-1/2 STORY – UNFINISHED ALL AGES
    e. 50 1-1/2 STORY FINISHED ALL AGES
    f. 60 2-STORY 1946 & NEWER
    g. 70 2-STORY 1945 & OLDER
    h. 75 2-1/2 STORY ALL AGES
    i. 80 SPLIT OR MULTI-LEVEL
    j. 85 SPLIT FOYERS
    k. 90 DUPLEX – ALL STYLES AND AGES
    l. 120 1-STORY PUD (Planned Unit Development) – 1946 & NEWER
    m. 150 1-1/2 STORY PUD – ALL AGES
    n. 160 2-STORY PUD – 1946 & NEWER
    o. 180 PUD – MULTILEVEL – INCL SPLIT LEV/FOYER
    p. 190 2 FAMILY CONVERSION – ALL STYLES AND AGES
2. Zone_Class: Identifies the general zoning classification of the sale
    a. A Agriculture
    b. C Commercial
    c. FV Floating Village Residential
    d. I Industrial
    e. RH Residential High Density

    f. RL Residential Low Density
    g. RP Residential Low-Density Park
    h. RM Residential Medium Density
3. LotFrontage: Linear feet of street-connected to the property
4. LotArea: Lot size is the lot or parcel side where it adjoins a street, boulevard or access way
5. Road_Type: Type of road access to the property
    a. Grvl Gravel
    b. Pave Paved
6. Alley: Type of alley access to the property
    a. Grvl Gravel
    b. Pave Paved
    c. NA No alley access
7. Property_Shape: General shape of the property
    a. Reg Regular
    b. IR1 Slightly irregular
    c. IR2 Moderately Irregular
    d. IR3 Irregular
8. LandContour: Flatness of the property
    a. Lvl Near Flat/Level
    b. Bnk Banked – Quick and significant rise from street grade to building
    c. HLS Hillside – Significant slope from side to side
    d. Low Depression
9. Utilities: Type of utilities available
    a. AllPub All public Utilities (E, G, W and S)
    b. NoSewr Electricity, Gas, and Water (Septic Tank)
    c. NoSeWa Electricity and Gas Only
    d. ELO Electricity only
10. LotConfig: Lot configuration
    a. Inside Inside lot
    b. Corner Corner lot
    c. CulDSac Cul-de-sac
    d. FR2 Frontage on 2 sides of property
    e. FR3 Frontage on 3 sides of property
11. LandSlope: Slope of property
    a. Gtl Gentle slope

b. Mod Moderate Slope
c. Sev Severe Slope

12. Neighborhood: Physical locations within Ames city limits
    a. Blmngtn Bloonington Heights
    b. Blueste Bluestem
    c. BrDale Briardale
    d. BrkSide Brookside
    e. ClearCr Clear Creek
    f. CollgCr College Creek
    g. Crawfor Crawford
    h. Edwards Edwards
    i. Gilbert Gilbert
    j. IDOTRR Iowa DOT and Rail Road
    k. MeadowV Meadow Village
    l. Mitchel Mitchell
    m. Names North Ames
    n. NoRidge Northridge
    o. NPkVill Northpark Villa
    p. NridgHt Northridge Heights
    q. NWAmes Northwest Ames
    r. OldTown Old Town
    s. SWISU South & West of Iowa State University
    t. Sawyer Sawyer
    u. SawyerW Sawyer West
    v. Somerst Somerset
    w. StoneBr Stone Brook
    x. Timber Timberland
    y. Veenker Veenker
13. Condition1: Proximity to various conditions
    a. Artery Adjacent to an arterial street
    b. Feedr Adjacent to feeder street
    c. Norm Normal
    d. RRNn Within 200’ of North-South Railroad
    e. RRAn Adjacent to North-South Railroad
    f. PosN Near positive off-site feature—park, greenbelt, etc.
    g. PosA Adjacent to positive off-site feature
    h. RRNe Within 200’ of East-West Railroad

i. RRAe Adjacent to East-West Railroad

14. Condition2: Proximity to various conditions (if more than one is present)
    a. Artery Adjacent to an arterial street
    b. Feedr Adjacent to feeder street
    c. Norm Normal
    d. RRNn Within 200’ of North-South Railroad
    e. RRAn Adjacent to North-South Railroad
    f. PosN Near positive off-site feature—park, greenbelt, etc.
    g. PosA Adjacent to positive off-site feature
    h. RRNe Within 200’ of East-West Railroad
    i. RRAe Adjacent to East-West Railroad
15. Dwelling_Type: Type of dwelling
    a. 1Fam Single-family Detached
    b. 2FmCon Two-family Conversion; originally built as a one-family dwelling
    c. Duplx Duplex
    d. TwnhsE Townhouse End Unit
    e. TwnhsI Townhouse Inside Unit
16. HouseStyle: Style of dwelling
    a. 1Story One story
    b. 1.5Fin One and one-half story: 2nd level finished
    c. 1.5Unf One and one-half story: 2nd level unfinished
    d. 2Story Two-story
    e. 2.5Fin Two and one-half story: 2nd level finished
    f. 2.5Unf Two and one-half story: 2nd level unfinished
    g. SFoyer Split Foyer
    h. SLvl Split Level
17. OverallQual: Rates the overall material and finish of the house
    a. 10 Very Excellent
    b. 9 Excellent
    c. 8 Very Good
    d. 7 Good
    e. 6 Above Average
    f. 5 Average
    g. 4 Below Average
    h. 3 Fair
    i. 2 Poor

j. 1 Very Poor

18. OverallCond: Rates the overall condition of the house
    a. 10 Very Excellent
    b. 9 Excellent
    c. 8 Very Good
    d. 7 Good
    e. 6 Above Average
    f. 5 Average
    g. 4 Below Average
    h. 3 Fair
    i. 2 Poor
    j. 1 Very Poor
19. YearBuilt: Original construction date
20. YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)
21. RoofStyle: Type of roof
    a. Flat Flat
    b. Gable Gble
    c. Gambrel Gabrel (Barn)
    d. Hip Hip
    e. Mansard Mansard
    f. Shed Shed
22. RoofMatl: Roof material
    a. ClyTile Clay or Tile
    b. CompShg Standard (Composite) Shingle
    c. Membran Membrane
    d. Metal Metal
    e. Roll Roll
    f. Tar&Grv Gravel & Tar
    g. WdShake Wood Shakes
    h. WdShngl Wood Shingles
23. Exterior1st: Exterior covering on the house
    a. AsbShng Asbestos Shingles
    b. AsphShn Asphalt Shingles
    c. BrkComm Brick Common
    d. BrkFace Brick Common
    e. CBlock Cinder Block

f. CemntBd Cement Board
g. HdBoard Hard Board
h. ImStucc Imitation Stucco
i. MetalSd Metal Siding
j. Other Other
k. Plywood Plywood
l. PreCast PreCast
m. Stone Stone
n. Stucco Stucco
o. VinylSd Vinyl Siding
p. WdSdng Wood Siding
q. WdShing Wood Shingles

24. Exterior2nd: Exterior covering on the house (if more than one material)
    a. AsbShng Asbestos Shingles
    b. AsphShn Asphalt Shingles
    c. BrkComm Brick Common
    d. BrkFace Brick Common
    e. CBlock Cinder Block
    f. CemntBd Cement Board
    g. HdBoard Hard Board
    h. ImStucc Imitation Stucco
    i. MetalSd Metal Siding
    j. Other Other
    k. Plywood Plywood
    l. PreCast PreCast
    m. Stone Stone
    n. Stucco Stucco
    o. VinylSd Vinyl Siding
    p. WdSdng Wood Siding
    q. WdShing Wood Shingles
25. MasVnrType: Masonry veneer type
    a. BrkCmn Brick Common
    b. BrkFace Brick Face
    c. CBlock Cider Block
    d. None None
    e. Stone Stone
26. MasVnrArea: Masonry veneer area in square feet

27. ExterQual: Evaluates the quality of the material on the exterior
    a. Ex Excellent
    b. Gd Good
    c. TA Average/Typical
    d. Fa Fair
    e. Po Poor
28. ExterCond: Evaluates the present condition of the material on the exterior
    a. Ex Excellent
    b. Gd Good
    c. TA Average/Typical
    d. Fa Fair
    e. Po Poor
29. Foundation: Type of foundation
    a. BrkTil Brick & Tile
    b. CBlock Cinder Block
    c. PConc Poured Concrete
    d. Slab Slab
    e. Stone Stone
    f. Wood Wood
30. BsmtQual: Evaluates the height of the basement
    a. Ex Excellent (100+ inches)
    b. Gd Good (90-99 inches)
    c. TA Typical (80-89 inches)
    d. Fa Fair (70-79 inches)
    e. Po Poor (<70 inches)
    f. NA No Basement
31. BsmtCond: Evaluates the general condition of the basement
    a. Ex Excellent
    b. Gd Good
    c. TA Typical – slight dampness allowed
    d. Fa Fair – dampness or some cracking or settling
    e. Po Poor – Severe cracking, settling, or wetness
    f. NA No Basement
32. BsmtExposure: Refers to walkout or garden level walls
    a. Gd Good Exposure

b. Av Average Exposure (split levels or foyers typically score average or above)
c. Mn Minimum Exposure
d. No No Exposure
e. NA No Basement

33. BsmtFinType1: Rating of basement finished area
    a. GLQ Good Living Quarters
    b. ALQ Average Living Quarters
    c. BLQ Below Average Living Quarters
    d. Rec Average Rec Room
    e. LwQ Low Quality
    f. Unf Unfinished
    g. NA No Basement
34. BsmtFinSF1: Type 1 finished square feet
35. BsmtFinType2: Rating of basement finished area (if multiple types)
    a. GLQ Good Living Quarters
    b. ALQ Average Living Quarters
    c. BLQ Below Average Living Quarters
    d. Rec Average Rec Room
    e. LwQ Low Quality
    f. Unf Unfinished
    g. NA No Basement
36. BsmtFinSF2: Type 2 finished square feet
37. BsmtUnfSF: Unfinished square feet of the basement area
38. TotalBsmtSF: Total square feet of the basement area
39. Heating: Type of heating
    a. Floor Floor Furnace
    b. GasA Gas forced warm air furnace
    c. GasW Gas hot water or steam heat
    d. Grav Gravity furnace
    e. OthW Hot water or steam heat other than gas
    f. Wall Wall furnace
40. HeatingQC: Heating quality and condition
    a. Ex Excellent
    b. Gd Good
    c. TA Average/Typical
    d. Fa Fair

    e. Po Poor
41. CentralAir: Central air conditioning
    a. N No
    b. Y Yes
42. Electrical: Electrical system
    a. SBrkr Standard Circuit Breaker & Romex
    b. FuseA Fuse Box over 60 AMP and all Romex wiring (Average)
    c. FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)
    d. FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)
    e. Mix Mixed
43. 1stFlrSF: First Floor square feet
44. 2ndFlrSF: Second floor square feet
45. LowQualFinSF: Low quality finished square feet (all floors)
46. GrLivArea: Above grade (ground) living area square feet
47. BsmtFullBath: Basement full bathrooms
48. BsmtHalfBath: Basement half bathrooms
49. FullBath: Full bathrooms above grade
50. HalfBath: Half bath above grade
51. Bedroom: Bedrooms above grade (does NOT include basement bedrooms)
52. Kitchen: Kitchens above grade
53. KitchenQual: Kitchen quality
    a. Ex Excellent
    b. Gd Good
    c. TA Average/Typical
    d. Fa Fair
    e. Po Poor
54. TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
55. Functional: Home functionality (Assume typical unless deduction are warranted)
    a. Typ Typical Functionality
    b. Min1 Minor Deductions 1
    c. Min2 Minor Deductions 2
    d. Mod Moderate Deductions
    e. Maj1 Major Deductions 1
    f. Maj2 Major Deductions2
    g. Sev Severely Damaged

h. Sal Salvage only

56. Fireplaces: Number of fireplaces
57. FireplaceQu: Fireplace quality
    a. Ex Excellent-Exceptional Masonry Fireplace
    b. Gd Good-Masonry Fireplace in the main level
    c. TA Average-Prefabricated Fireplace in the main living area or Masonry Fireplace in basement
    d. Fa Fair-Prefabricated Fireplace in a basement
    e. Po Poor-Ben Franklin Stove
    f. NA No Fireplace
58. GarageType: Garage location
    a. 2Types More than one type of garage
    b. Attchd Attached to the home
    c. Basment Basement Garage
    d. BuiltIn Built-In (Garage part of the house-typically has room above garage)
    e. CarPort Car Port
    f. Detchd Detached from home
    g. NA No Garage
59. GarageYrBlt: Year garage was built
60. GarageFinish: Interior finish of the garage
    a. Fin Finished
    b. RFn Rough Finished
    c. Unf Unfinished
    d. NA No Garage
61. GarageCars: Size of garage in car capacity
62. GarageArea: Size of garage in square feet
63. GarageQual: Garage quality
    a. Ex Excellent
    b. Gd Good
    c. TA Average/Typical
    d. Fa Fair
    e. Po Poor
    f. NA No Garage
64. GarageCond: Garage condition
    a. Ex Excellent
    b. Gd Good

    c. TA Average/Typical
    d. Fa Fair
    e. Po Poor
    f. NA No Garage
65. PavedDrive: Paved driveway
    a. Y Paved
    b. P Partial Pavement
    c. N Dirt/Gravel
66. WoodDeckSF: Wood deck area in square feet
67. OpenPorchSF: Open porch area in square feet
68. EnclosedPorch: Enclosed porch area in square feet
69. 3SsnPorch: Three season porch area in square feet
70. ScreenPorch: Screen porch area in square feet
71. PoolArea: Pool area in square feet
72. PoolQC: Pool quality
    a. Ex Excellent
    b. Gd Good
    c. TA Average/Typical
    d. Fa Fair
    e. NA No Pool
73. Fence: Fence quality
    a. GdPrv Good Privacy
    b. MnPrv Minimum Privacy
    c. GdWo Good Wood
    d. MnWw Minimum Wood/Wire
    e. NA No Fence
74. MiscFeature: Miscellaneous feature not covered in other categories
    a. Elev Elevator
    b. Gar2 2nd Garage (if not described in garage section)
    c. Othr Other
    d. Shed Sherd (over 100 SF)
    e. TenC Tennis Court
    f. NA None
75. MiscVal: Value of miscellaneous feature
76. MoSold: Month Sold (MM)
77. YrSold: Year Sold (YYYY)
78. SaleType: Type of sale

a. WD Warranty Deed-Conventional
b. CWD Warranty Deed-Cash
c. VWD Warrenty Deed-VA Loan
d. New Home just constructed and sold
e. COD Court Officer Deed/Estate
f. Con Contract 15% Down payment regular terms
g. ConLw Contract Low Down payment and low interest
h. ConLI Contract Low Interest
i. ConLD Contract Low Down
j. Oth Other

79. SaleCondition: Condition of sale
    a. Normal Normal Sale
    b. Abnormal Abnormal Sale – trade, foreclosure, short sale
    c. AdjLand Adjoining Land Purchase
    d. Alloca Allocation – two linked properties with separate deeds, typically condo with a garage unit
    e. Family Sale between family members
    f. Partial Home was not completed when last assessed (associated with New Homes)
80. Property_Sale_Price: Price of the house

4. Understanding the data

4.1. Missing Values

The following columns contain missing values:

- PoolQC (2065 missing data points)
- MiscFeature (1993 missing data points)
- Alley (1944 missing data points)
- Fence (1669 missing data points)
- FireplaceQu (988 missing data points)
- LotFrontage (257 missing data points)
- MasVnrType (10 missing data points)
- Mas VnrArea (10 missing data points)
- GarageYrBlt (90 missing data points)

If the missing values are above 40% of the data, then I removed that column. Hence the removed columns are:

- PoolQC (99.61% of missing values)
- MiscFeature (96.14% of missing values)
- Alley (93.78% of missing values)
- Fence (80.51% of missing values)
- FireplaceQu (47.66% of missing values)

The remaining columns having missing values are filled with mean or median of the column if the column is continuous, and mode of the column if the column is categorical.

4.2 Response Variable

The response variable or dependent variable of this data is Property Sale Price. It is the price at which the property sold off. It is a continuous variable with minimum price of $ 35,311 and maximum price of $ 755,000. The mean price of property is $ 182,290 with a standard deviation of $ 79,156.

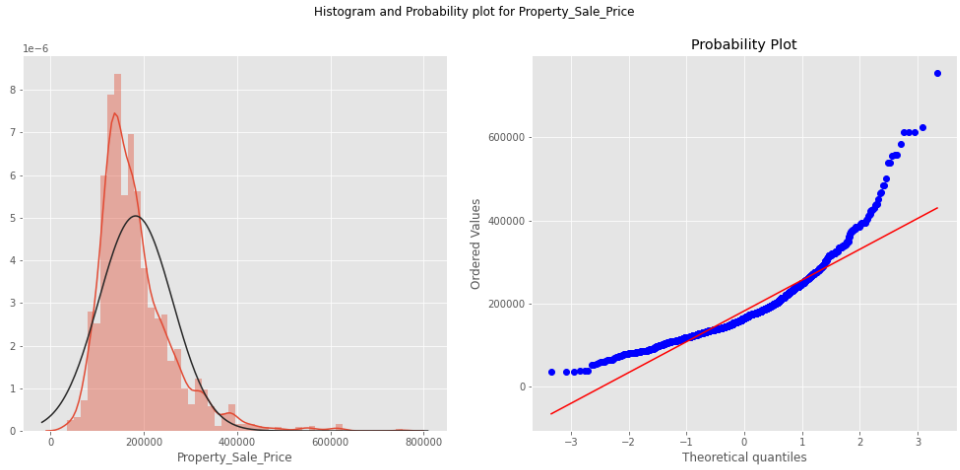

Figure 1: Histogram and QQ-plot of Property Sale Price

Figure 1 shows the Histogram and QQ-plot of property sale price. The Shapiro-Wilk test for normality done on property sale price yield a p-value of 0, hence I rejected the null hypothesis and accepted the alternate hypothesis which states: 'There is a significant departure from normal bell-shaped curve (Data is not normal)'. As Linear regression assumes the data to be normally distributed, I transformed the column to a normal distribution using Power Transformer function provided by Scikit-learn library.

The box plot of Property sale price (*Figure 2*) shows the presence of outliers in the column. There are 68 outliers in the column, which should be removed before training the model.

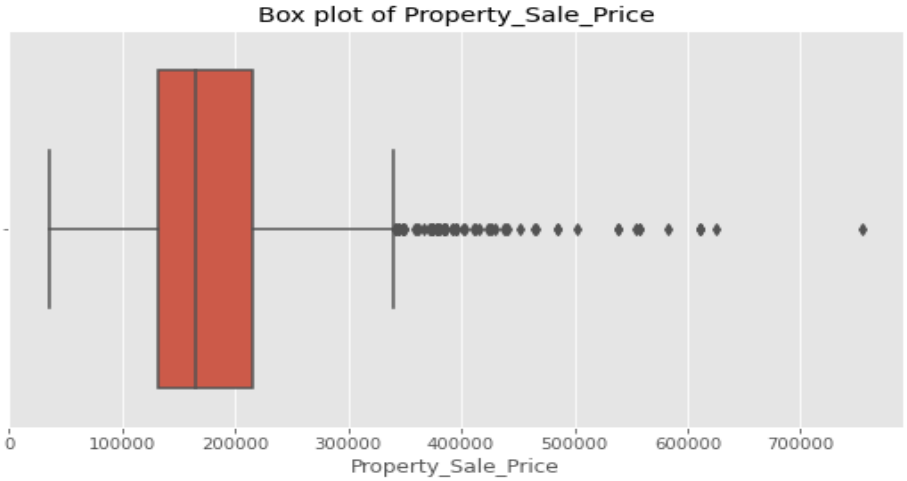

Figure 2: Boxplot of Property Sale Price

4.3. Continuous Variables

Figure 3 shows the heatmap of all the continuous variables including the variables representing year. From the heatmap I found out that there is an issue of multicollinearity with some columns. Multicollinearity is the correlation of independent variables with each other. A Linear regression model assumes that there is no multicollinearity in the data. The best measure to find multicollinearity in data is to find the Variance Inflation Factor (VIF). If VIF factor is equal to 1, then there is no correlation. If VIF factor is between 1 and 5, then the variables are moderately correlated. If VIF factor is more than 5, then the variables are highly correlated.

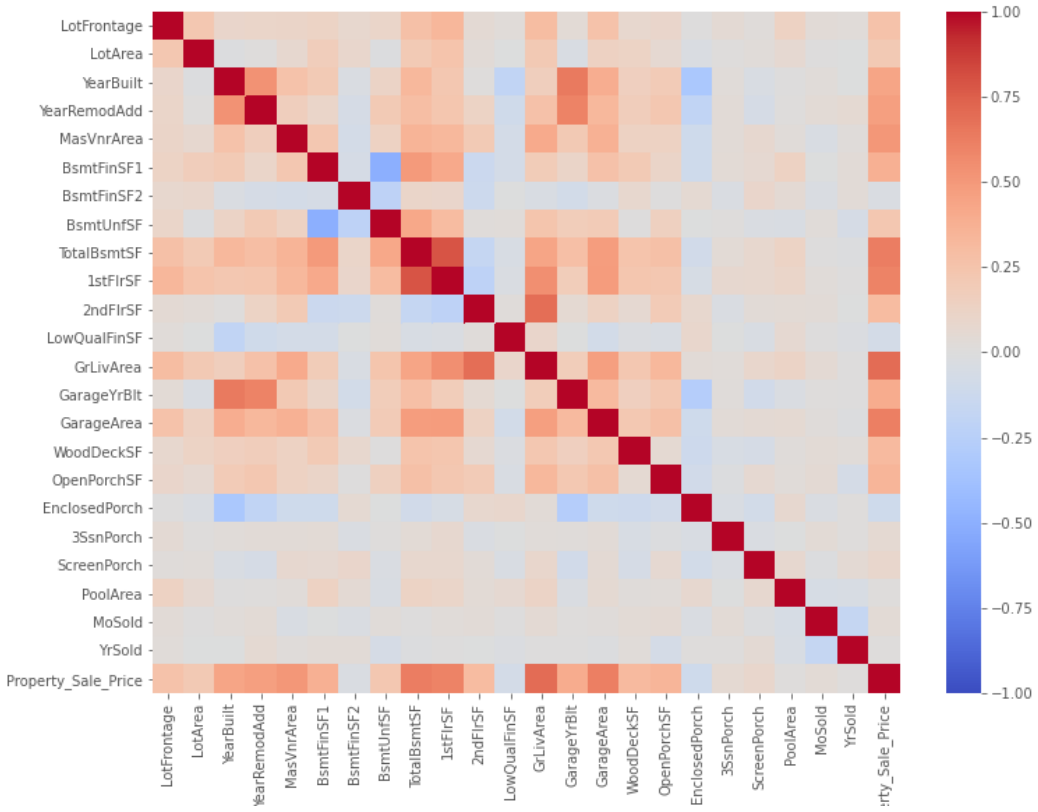

Figure 3: Heatmap of correlation matrix (before VIF)

After conducting VIF factor analysis, I had to remove the following columns on the basis of multicollinearity:

- BsmtFinSF1
- 1stFlrSF
- YearRemodAdd
- GarageYrBlt
- YrSold
- GarageArea
- LotFrontage
- TotalBsmtSF

After removing these columns correlation between independent variables decreased and the issue of multicollinearity between continuous variables is solved.

Figure 4 shows the heatmap of correlation matrix of variables after the VIF process and removal of variables with high correlation.

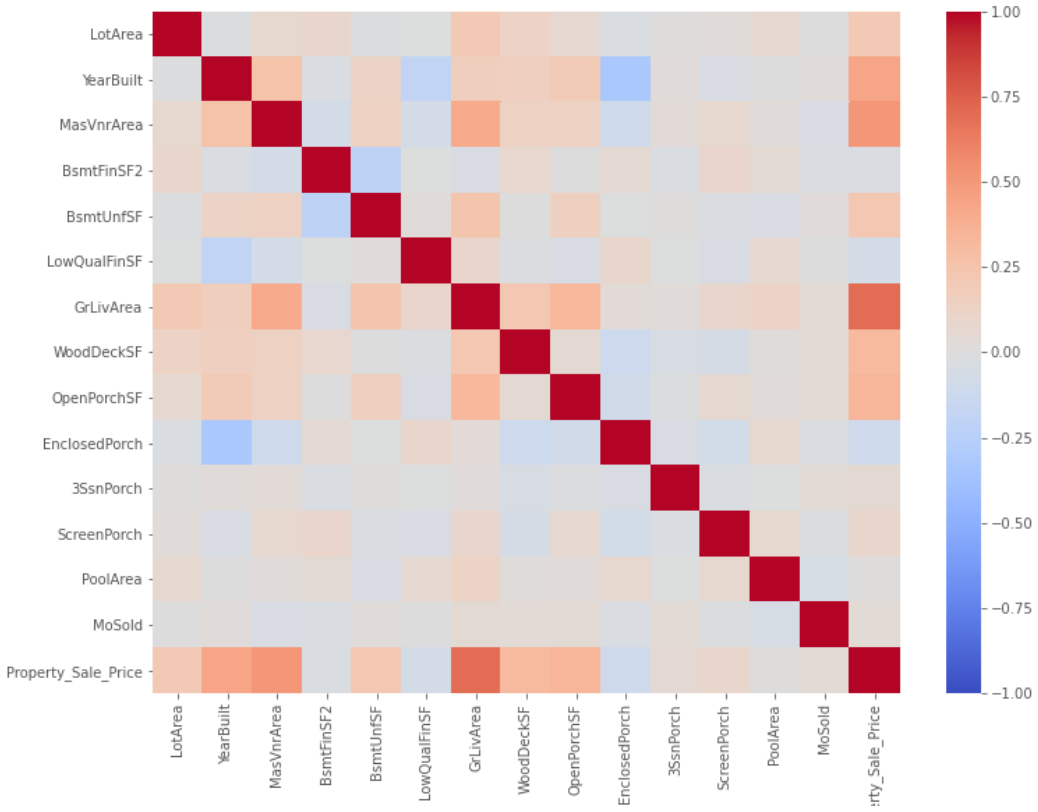

Figure 4: Heatmap of correlation matrix (after VIF)

4.4. Categorical variables

There are more categorical variables than continuous variable in the dataset. All the categorical variables which are used for the Model should be encoded into 1's and 0's for the algorithm to learn. This will create more columns and increase the complexity of the model. Also, some of the levels in the categorical variables have very less proportion compared to other levels. To avoid this, I clubbed some levels together.

The levels of Dwell_Type is clubbed into 2, which are '20' and '60', the levels of Zone_Class is clubbed into 2, which are 'low_density' and 'medium_density'. Similarly for all other categorical variables, if it is required then this process is done.

4.5. p-value

For all the variables Hypothesis testing is done and the p-value is noted. If the p-value is below alpha (which is taken as 0.05) then we reject the null hypothesis and the variable is a good predictor of property sale price. If the p-

value is above alpha then we fail to reject null hypothesis and the variable is not a good predictor of property sale price.

column	p-value	column	p-value
Id	0.927664303	HeatingQC	4.61E-92
Dwell_Type	5.58E-08	CentralAir	7.97E-29
Zone_Class	3.54E-25	Electrical	6.22E-24
LotFrontage	9.74E-31	1stFlrSF	3.72E-160
LotArea	2.29E-53	2ndFlrSF	7.37E-42
Road_Type	0.076648796	TotalSF	3.36E-288
Property_Shape	4.08E-32	LowQualFinSF	0.000662509
LandContour	2.09E-12	GrLivArea	2.60E-256
Utilities	0.288151956	BsmtFullBath	6.17E-21
LotConfig	9.39E-05	BsmtHalfBath	0.235037206
LandSlope	0.120331011	FullBath	3.72E-150
Neighborhood	1.40E-18	HalfBath	4.37E-39
Condition1	2.14E-12	TotalBath	2.02E-145
Condition2	0.007706426	BedroomAbvGr	4.32E-16
Dwelling_Type	3.11E-17	KitchenAbvGr	8.61E-10
HouseStyle	2.33E-14	KitchenQual	4.62E-148
OverallQual	5.48E-137	TotRmsAbvGrd	2.88E-89
OverallCond	8.04E-27	Functional	4.23E-06
YearBuilt	3.07E-93	Fireplaces	5.09E-112
AgeOfHouse	3.07E-93	GarageType	1.63E-68
YearRemodAdd	2.68E-109	GarageYrBlt	1.55E-72
RoofStyle	2.91E-16	GarageFinish	7.56E-158
RoofMatl	5.45E-06	GarageCars	2.50E-249
Exterior1st	6.61E-56	GarageArea	1.24E-204
Exterior2nd	3.68E-53	GarageQual	7.13E-10
MasVnrType	1.63E-62	GarageCond	1.89E-09
MasVnrArea	1.59E-89	PavedDrive	7.02E-20
ExterQual	5.01E-238	WoodDeckSF	2.49E-39
ExterCond	3.30E-06	OpenPorchSF	4.98E-72
Foundation	2.82E-129	EnclosedPorch	5.92E-06
BsmtQual	1.33E-122	3SsnPorch	0.011303838
BsmtCond	1.18E-13	ScreenPorch	0.000464631
BsmtExposure	4.06E-33	PoolArea	0.515383977
BsmtFinType1	1.78E-85	MiscVal	0.602538479
BsmtFinSF1	1.97E-56	MoSold	0.008860919
BsmtFinType2	0.011517832	YrSold	0.932498304
BsmtFinSF2	0.364785174	SaleType	1.45E-37
BsmtUnfSF	3.45E-23	SaleCondition	7.36E-42
TotalBsmtSF	5.98E-203		
Heating	2.30E-07		

Table 1: p-values of all columns

Based on the p-values the following columns have no significant relation/impact on the response variable which is property sale price.

- 2ndFlrSF
- 3SsnPorch
- BsmtFinSF2
- BsmtFullBath
- BsmtHalfBath
- Dwell_Type
- EnclosedPorch
- FullBath
- HalfBath
- Id
- LandSlope
- LowQualFinSF
- MiscVal
- PoolArea
- Road_Type
- ScreenPorch
- Utilities
- YearBuilt

4.6. Proportion

If the levels of a categorical variable have uneven proportion, where some of the levels dominate other levels, and clubbing the levels won't solve this problem, then that column has to be dropped. Due to this problem the following columns are removed from the final data.

- BsmtCond
- BsmtFinType2
- CentralAir
- Condition1
- Condition2
- Dwelling_Type
- Electrical
- ExterCond
- Functional

- GarageCond
- GarageQual
- Heating
- KitchenAbvGr
- LandContour
- PavedDrive
- RoofMatl
- RoofStyle
- SaleCondition
- SaleType
- Zone_Class

4.7. Conclusion

After the Exploratory Data Analysis, all the unnecessary variables are removed from the dataset and only the necessary variables are used for model building. From 80 independent variables I selected only 28 variables for my final model.

The selected variables are:

1. LotArea
2. Property_Shape
3. LotConfig
4. Neighborhood
5. HouseStyle
6. OverallQual
7. OverallCond
8. Exterior1st
9. Exterior2nd
10. MasVnrType
11. MasVnrArea
12. ExterQual
13. Foundation
14. BsmtQual
15. BsmtExposure
16. BsmtFinType1
17. BsmtUnfSF
18. HeatingQC
19. GrLivArea

20. BedroomAbvGr
21. KitchenQual
22. TotRmsAbvGrd
23. Fireplaces
24. GarageType
25. GarageFinish
26. GarageCars
27. WoodDeckSF
28. OpenPorchSF

5. Model Analysis

5.1. Linear Regression Model 1 (Full Model)

Initially I build a Linear regression model with all the variables as a base model. The idea behind a base model is for the comparison with all the other model.

Before building the base model, all the numerical variables are Scaled using StandardScaler function provided by Scikit-learn library. The StandardScaler will scale all the numerical variables to a comfortable scale. Linear regression model uses ordinary least square (OLS) method to find the best fit line which is the predicted values. To find the best fit line, sum of squared errors is calculated. Since the model is calculating Euclidian distance, all the variables need to be scaled to value around 0. Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data (e.g., Gaussian with 0 mean and unit variance).

In practice we often ignore the shape of the distribution and just transform the data to center it by removing the mean value of each feature, then scale it by dividing non-constant features by their standard deviation.

In case of categorical variables, encoding should be done. A linear regression model cannot process string variables. All the non-numerical variables should be converted to numbers. Here OneHotEncoder is used to convert non-numerical variables to numbers. OneHotEncoder is a function provided by Scikit-learn library. The input to this transformer should be an array-like of integers or strings, denoting the values taken on by categorical (discrete) features. The features are encoded using a one-hot (aka ‘one-of-K’ or ‘dummy’)

encoding scheme. This creates a binary column for each category and returns a sparse matrix or dense array (depending on the sparse parameter)

The data should be split into training set and testing set. The training set is used for the training of the model and the testing set is used to test the model on new data points and the error is calculated. The test size selected is 20 percent of the whole data. Learning the parameters of a prediction function and testing it on the same data is a methodological mistake: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data. This situation is called overfitting. To avoid it, it is common practice when performing a (supervised) machine learning experiment to hold out part of the available data as a test set X_test, y_test. Note that the word “experiment” is not intended to denote academic use only, because even in commercial settings machine learning usually starts out experimentally. Here is a flowchart of typical cross validation workflow in model training. The best parameters can be determined by grid search techniques.

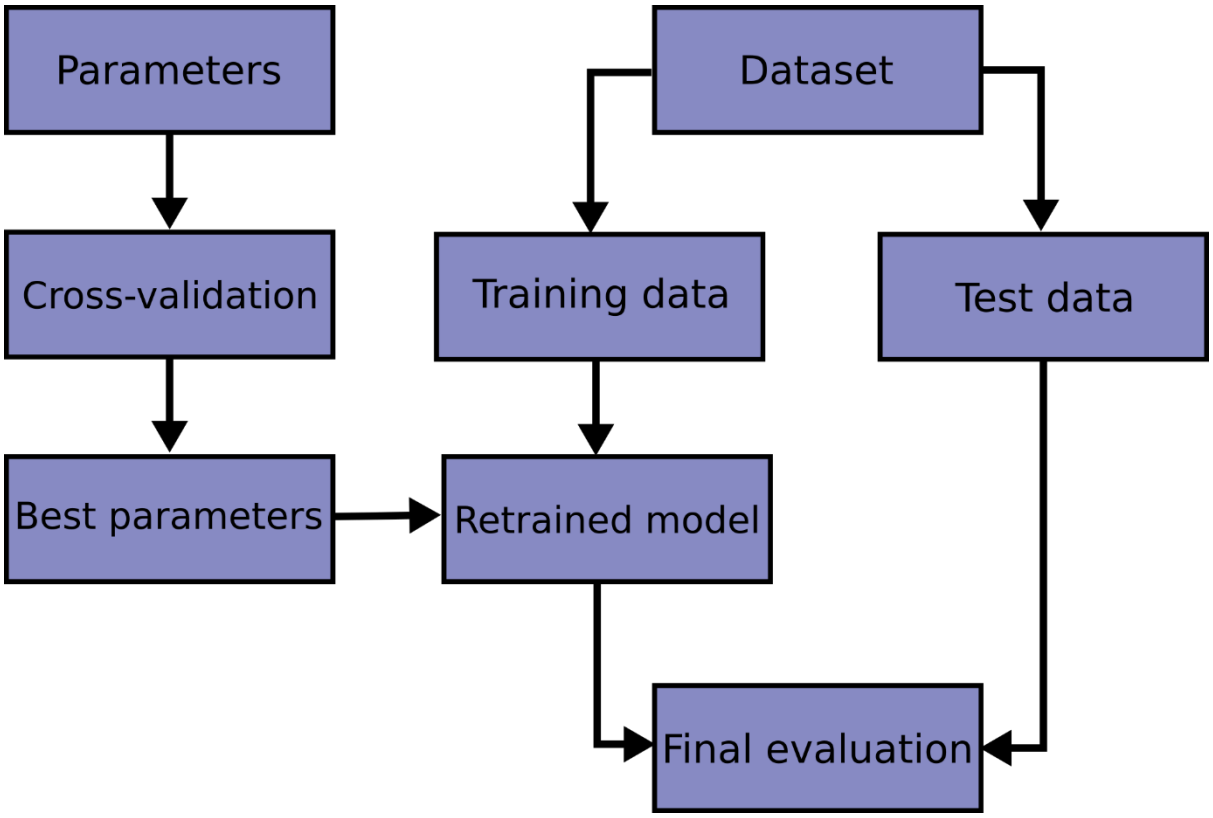

Figure 5: Grid search workflow

A Linear model is trained using the training set and predictions are made on the testing set. The LinearRegression class of Scikit-learn library is used to build the model.

After predicting the values on testing set, errors are calculated and a histogram of errors are plotted, which is shown in *Figure 5.*

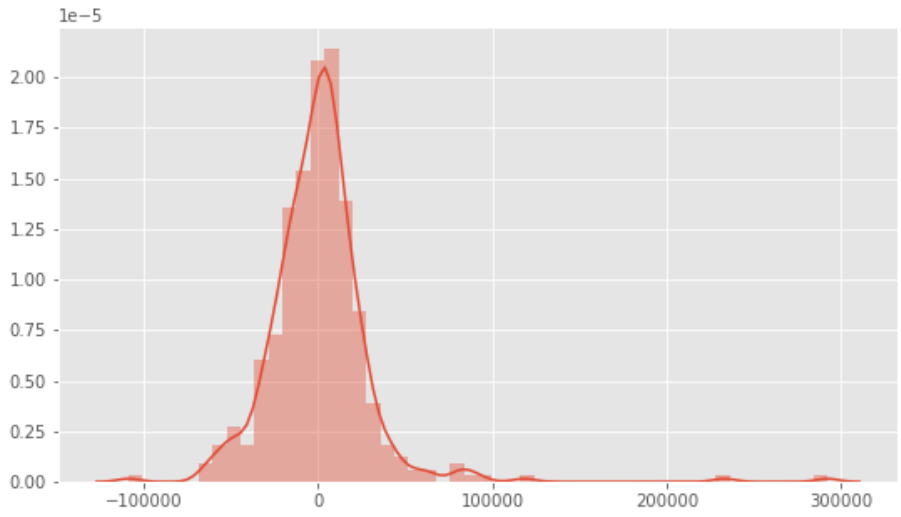

Figure 6: Histogram of errors

The metrics used to calculate the errors are Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

Mean Absolute Error (MAE): In statistics, mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon. Examples of Y versus X include comparisons of predicted versus observed, subsequent time versus initial time, and one technique of measurement versus an alternative technique of measurement. It has the same unit as the original data, and it can only be compared between models whose errors are measured in the same units. It is usually similar in magnitude to RMSE, but slightly smaller. MAE is calculated as:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

It is thus an arithmetic average of the absolute errors, where yi is the prediction and xi the actual value. Note that alternative formulations may include relative frequencies as weight factors. The mean absolute error uses

the same scale as the data being measured. This is known as a scale-dependent accuracy measure and, therefore cannot be used to make comparisons between series using different scales.

Mean Squared Error (MSE): Mean Squared Error (MSE) or Mean Squared Deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors — that is, the average squared difference between the estimated values and the actual value. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate.

The MSE assesses the quality of a predictor (i.e., a function mapping arbitrary inputs to a sample of values of some random variable), or an estimator (i.e., a mathematical function mapping a sample of data to an estimate of a parameter of the population from which the data is sampled). The definition of an MSE differs according to whether one is describing a predictor or an estimator.

The MSE is a measure of the quality of an estimator — it is always non-negative, and values closer to zero are better.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Let's analyze what this equation actually means.

- In mathematics, the character that looks like weird E is called summation (Greek sigma). It is the sum of a sequence of numbers, from i=1 to n. Let's imagine this like an array of points, where we go through all the points, from the first (i=1) to the last (i=n).
- For each point, we take the y-coordinate of the point, and the y'-coordinate. We subtract the y-coordinate value from the y'-coordinate value and calculate the square of the result.
- The third part is to take the sum of all the $(y-y')^2$ values and divide it by n, which will give the mean.

Our goal is to minimize this mean, which will provide us with the best line that goes through all the points.

Root Mean Squared Error on Prediction (RMSE / RMSEP): In statistical modeling and particularly regression analyses, a common way of measuring the quality of the fit of the model is the RMSE (also called Root Mean Square Deviation), given by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n}}$$

where yi is the ith observation of y and ŷ the predicted y value given the model. If the predicted responses are very close to the true responses the RMSE will be small. If the predicted and true responses differ substantially — at least for some observations — the RMSE will be large. A value of zero would indicate a perfect fit to the data. Since the RMSE is measured on the same scale, with the same units as y, one can expect 68% of the y values to be within 1 RMSE — given the data is normally distributed.

The MAE, MSE and RMSE of the data is given below.

Mean Absolute Error (MAE)	18867.86303088855
Mean Squared Error (MSE)	938322583.5631702
Root Mean Squared Error (RMSE)	30632.051572873308

Coefficient of Determination (R2): R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, if the R2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

The r2-score of linear model-1 is 0.8398414871691536, which means that 83.98 % of variance in Property sale price is explained by the independent variables collectively.

5.2. Linear Regression Model 2

A new Linear regression model is created with only the selected variables. I have selected 28 independent variables which are necessary for predicting Property sale price.

The same process of Scaling and Encoding should be done to the data input for this model. After this process the model is trained on training set and then property sale price is predicted on testing set.

After predicting the values, errors are calculated using the formula y_test – y_predicted. *Figure 6* shows the histogram of errors.

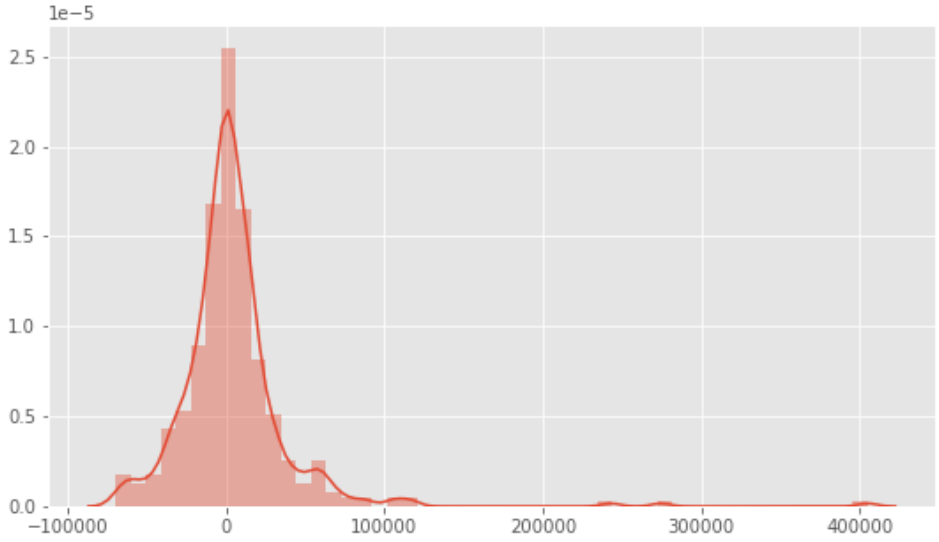

Figure 7: Histogram of errors

The metrics used to calculate the errors are Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

Mean Absolute Error (MAE)	20908.481744555247
Mean Squared Error (MSE)	1459913381.602235
Root Mean Squared Error (RMSE)	38208.81287873565

The r2-score of linear model-2 is 0.7508132489240843, which means that 75.08 % of variance in Property sale price is explained by the independent variables collectively.

6. Advantages and Limitations

6.1. Advantages of the model

- A Linear regression model is simple to implement and easier to interpret the output coefficients.
- When you know the relationship between the independent and dependent variable have a linear relationship, this algorithm is the best to use because of its less complexity when compared to other algorithms.
- Linear regression is susceptible to over-fitting but it can be avoided using some dimensionality reduction techniques like regularization (l1 and l2) techniques and cross-validation.
- The modeling of the predictions as a weighted sum makes it transparent as to how predictions are produced. And with Lasso we can ensure that the number of features used remains small.
- Many people use linear regression models. This means that in many places it is accepted for predictive modeling and doing inference. There is a high level of collective experience and expertise, including teaching materials on linear regression models and software implementations.
- Mathematically, it is straightforward to estimate the weights and you have a guarantee to find optimal weights (given all assumptions of the linear regression model are met by the data.)
- Together with the weights you get confidence intervals, tests, and solid statistical theory.
- The final model (Linear regression model 2) has only 28 variables.

6.2. Limitations of the model

- In linear regression technique outliers can gave huge effects on the regression and boundaries are linear in this technique. Hence the outliers in all the column were removed for increasing the performance of the model.

- Diversely, linear regression assumes a linear relationship between dependent and independent variables. That means it assumes that there is a straight-line relationship between them. Each nonlinearity or interaction has to be hand-crafted and explicitly given to the model as an input feature. Also, it assumes independence between attributes.
- But then linear regression also looks at a relationship between the mean of the dependent variables and independent variables. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables.
- Before applying Linear regression, multicollinearity should be removed (using dimensionality reduction techniques) because it assumes that there is no relationship among independent variables.
- The interpretation of a weight can be unintuitive because it depends on all other features. A feature with high positive correlation with the outcome y and another feature might get a negative weight in the linear model because, given the other correlated feature, it is negatively correlated with y in the high-dimensional space.
- Completely correlated feature makes it even impossible to find a unique solution for the linear equation. House size and number of rooms are highly correlated: the bigger a house is, the more rooms it has. If you take both features into a linear model, it might happen, that the size of the house is the better predictor and gets a large positive weight. The number of rooms might end up getting a negative weight, because, given that a house has the same size, increasing the number of rooms could make it less valuable or the linear equation becomes less stable, when the correlation is too strong.

6.3. Summary

Linear regression is a great tool to analyze the relationships among the variables but it isn't recommended for most practical applications because it over-simplifies real-world problems by assuming a linear relationship among the variables. There are some assumptions (Linearity, Homoskedasticity, no-multicollinearity, no autocorrelation…) that should be considered about the data for a linear regression model.

7. Conclusion

The analysis started with 80 variables for predicting the sale price of a property. But due to missing values, uneven proportion, non-linearity, multicollinearity and based on p-values 52 variables were dropped and only 28 variables were considered for the final model.

The final model (Linear regression model 2) explains 75.08% of variance in property sale price. But when compared to the base model which explains 83.98% of variance, the final model is performing poorly. But the final model has very few variables which makes it a simple model, and all the assumptions of a linear regression were satisfied for the final model.

8. References

1. https://en.wikipedia.org/wiki/Linear_regression
2. http://www.holehouse.org/mlclass/04_Linear_Regression_with_multiple_variables.html
3. http://machinelearningmastery.com/simple-linear-regression-tutorial-for-machine-learning
4. https://ml-cheatsheet.readthedocs.io/en/latest/linear_regression.html#summary
5. https://www.analyticsvidhya.com/blog/2015/08/common-machine-learning-algorithms
6. Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. "The elements of statistical learning". www.web.stanford.edu/~hastie/ElemStatLearn/ (2009).
7. Ken Black, “Business for contemporary decision-making statistics”-Sixth edition. Unit 4 – Regression analysis and forecasting.
8. Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong, “Mathematics for Machine Learning”. Chapter 9 – Linear Regression.
9. Dr. EaswaranIyer, Dr. Vinod Kumar Murti, “Segmentation of Indian States based on select economic variables using k-means Non-Hierarchical Clustering”
10. https://towardsdatascience.com/ways-to-evaluate-regression-models-77a3ff45ba70
11. https://www.edupristine.com/blog/detecting-multicollinearity

12. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
13. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html
14. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
15. https://www.geeksforgeeks.org/ml-advantages-and-disadvantages-of-linear-regression/
16. https://christophm.github.io/interpretable-ml-book/limo.html
17. https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd