

**HOUSING- PRICE PREDICTION**.

Submitted by:

ABHISHEK GUPTA

**ACKNOWLEDGMENT**

I would like to acknowledge everyone who played a role in my academic Accomplishments. First of all, my guided teacher and employer who helped me in doing in this project. I have used my proper knowledge and guidance to work in this project at maximum capacity.

In this project I have include the research papers, various machine learning algorithms, websites which helped me to analyse the data and then take corrective measures in it and at last the given dataset from which I pulled out the relevant information.

INTRODUCTION

**Business Problem:**

In this project, A US-based housing company named **Surprise Housing** has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia.

The company is looking at prospective properties to buy houses to enter the market. You are required to build a model using Machine Learning to predict the actual value of the prospective properties and decide whether to invest in them or not. For this company wants to know:

Which variables are important to predict the price of variable?

How do these variables describe the price of the house?

The goal of business is to predict the model price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

**Conceptual Background of the Domain Problem:**

The main domain and conceptual problem of the given data that is provided to us is indicating various definitions of the data description where the concluding column or the predictive column will be analysed by applying various machine learning algorithms such as Linear Regression, Decision Tree Regressor, Random Forest Regressor, Extra Tree Regressor etc.

These data columns are the indications of a particular reason which helps the client to make a future prediction based on the data analysis of all the affecting factors that could make a difference in deciding the sale price of a house. The size of the train data is 1168 rows × 81 columns which describes the data records which are as follows.

Most of the data contains object values then integer values with having null values as can be seen in the heatmap.

This data description is useful for the better understanding of my project which will be calculated after analyzing the different columns and understanding their co-relation with the label column and overall efficiency of the given dataset.

**Review of Literature**

This is a comprehensive summary of my research done which is based on the various factors such as data cleaning, doing exploratory data analysis, finding out the null values and then correcting them with the next best alternative, detecting outliers, pre-processing the data, applying various machine learning algorithms and then finally deciding, implementing that algorithms which is giving the maximum accuracy with the least number of errors.

In the process of analysing the data I have used various methods and coding techniques to get the better understanding of the data.

**Motivation for the Problem Undertaken**

The objective of doing this project is to determine whether to decide the entry of surprise housing company in the Australian Market or not. If so then determine the future predictive analysis of the sale price of the houses based upon the factors which could affects the overall pricing of house such as location, year of build, what are the facilities available in the particular house and then decide whether to make an investment or not.

**Analytical Problem Framing**

**Mathematical/ Analytical Modeling of the Problem**

The mathematical analysis of the given data is understandable in the following way:

* There are too zero values in the dataset which contains many outliers.
* In all the data present the mean values are greater than the median values which indicates that our data is rightly skewed that is the data is not at all normally distributed.
* There is too much gap between the 75th percentile values and the maximum values which is the clear-cut indication of large no. of outliers that are present in our dataset.
* The data is very deviated from its mean value therefore indicates outliers are present.

**Data Sources and their formats:**

In the given data set the data description defines the basis of each column that ultimately affect the effectiveness and efficiencies of the label column as it indicates the success or the failure in terms of returning the credit amount with in the range period of 5 days.

The data description and their formats as follows which helps us to get a better understanding of the problem statement in analysing the given dataset.:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 unit | | | | | | | |
| Family | Sale between family members | | |  |  |  |  |  |
| Partial | Home was not completed when last assessed (associated with New Homes) | | | | | | | |
|  |  | | | | | | | |

The given columns describe the meaning of the data that belongs to the house or the factors which affects the determination of sale price for a particular house.

The information regarding the data types, values counts and the null values are as follows:

0 Id 1168 non-null int64

1 MSSubClass 1168 non-null int64

2 MSZoning 1168 non-null object

3 LotFrontage 954 non-null float64

4 LotArea 1168 non-null int64

5 Street 1168 non-null object

6 Alley 77 non-null object

7 LotShape 1168 non-null object

8 LandContour 1168 non-null object

9 Utilities 1168 non-null object

10 LotConfig 1168 non-null object

11 LandSlope 1168 non-null object

12 Neighborhood 1168 non-null object

13 Condition1 1168 non-null object

14 Condition2 1168 non-null object

15 BldgType 1168 non-null object

16 HouseStyle 1168 non-null object

17 OverallQual 1168 non-null int64

18 OverallCond 1168 non-null int64

19 YearBuilt 1168 non-null int64

20 YearRemodAdd 1168 non-null int64

21 RoofStyle 1168 non-null object

22 RoofMatl 1168 non-null object

23 Exterior1st 1168 non-null object

24 Exterior2nd 1168 non-null object

25 MasVnrType 1161 non-null object

26 MasVnrArea 1161 non-null float64

27 ExterQual 1168 non-null object

28 ExterCond 1168 non-null object

29 Foundation 1168 non-null object

30 BsmtQual 1138 non-null object

31 BsmtCond 1138 non-null object

32 BsmtExposure 1137 non-null object

33 BsmtFinType1 1138 non-null object

34 BsmtFinSF1 1168 non-null int64

35 BsmtFinType2 1137 non-null object

36 BsmtFinSF2 1168 non-null int64

37 BsmtUnfSF 1168 non-null int64

38 TotalBsmtSF 1168 non-null int64

39 Heating 1168 non-null object

40 HeatingQC 1168 non-null object

41 CentralAir 1168 non-null object

42 Electrical 1168 non-null object

43 1stFlrSF 1168 non-null int64

44 2ndFlrSF 1168 non-null int64

45 LowQualFinSF 1168 non-null int64

46 GrLivArea 1168 non-null int64

47 BsmtFullBath 1168 non-null int64

48 BsmtHalfBath 1168 non-null int64

49 FullBath 1168 non-null int64

50 HalfBath 1168 non-null int64

51 BedroomAbvGr 1168 non-null int64

52 KitchenAbvGr 1168 non-null int64

53 KitchenQual 1168 non-null object

54 TotRmsAbvGrd 1168 non-null int64

55 Functional 1168 non-null object

56 Fireplaces 1168 non-null int64

57 FireplaceQu 617 non-null object

58 GarageType 1104 non-null object

59 GarageYrBlt 1104 non-null float64

60 GarageFinish 1104 non-null object

61 GarageCars 1168 non-null int64

62 GarageArea 1168 non-null int64

63 GarageQual 1104 non-null object

64 GarageCond 1104 non-null object

65 PavedDrive 1168 non-null object

66 WoodDeckSF 1168 non-null int64

67 OpenPorchSF 1168 non-null int64

68 EnclosedPorch 1168 non-null int64

69 3SsnPorch 1168 non-null int64

70 ScreenPorch 1168 non-null int64

71 PoolArea 1168 non-null int64

72 PoolQC 7 non-null object

73 Fence 237 non-null object

74 MiscFeature 44 non-null object

75 MiscVal 1168 non-null int64

76 MoSold 1168 non-null int64

77 YrSold 1168 non-null int64

78 SaleType 1168 non-null object

79 SaleCondition 1168 non-null object

80 SalePrice 1168 non-null int64

dtypes: float64(3), int64(35), object(43)

**Data Preprocessing Done:**

In this section I have covered a part which includes the data cleaning, assumption made, modification and updating of the data to get the relevant insights or the information out of it. This process is necessary to get the more useful data which can be used, removes the unnecessary data which hinders the process of data visualizations and performing the model selection part where we can decide the future investment strategies based upon the affecting factors. The process is divided into various steps which are performed to put our data in normal distribution at possible.

The steps are as follows:

1. After checking the value counts of the columns Alley, PoolQC, Fence, Mics Feature, FireplaceQu it indicates that more than 80% of the data contains null values so in this aspect I have dropped all these columns as there is no sense of continue with the unrelated data for the given dataset. So to clean the data as per my understanding of the data I have drop the columns which are irrelevant to my data.
2. In the column Utilities the value counts indicates that there is only one data attribute which is affecting my dataset so in order to treat this problem I have drop this column as there is no unique points to be consider for the given column also it is a object column so we can drop this column as well as there is no use of this column in evaluating the data for the better understanding of our given model.
3. In the year bulit column the datatype is in object format so to analyse that column we must impute its value in integer datatype. By applying label encoder, we can distinguish between years of the build period After this it is easy to understand that at which point of time there is high selling price for the buildings.
4. Encoding the object columns to analyse at which the given dataset for predicting the sale price of the houses.
5. Scaling the data to get the data into standard form to perform the model selection process.
6. Detecting the outliers at which our data is having the extreme values by zscore method.

**Data Inputs- Logic- Output Relationships:**

In this section I have conclude that most of the data column are directly co-related to its output variable that is label column which helps in deciding the success or the failure in the investment of surprise housing company.

Chart, bar chart

Description automatically generated

If I look into the co-relation figure, we can see label column is dependent on the values of its input column except the column of KichenAbvGr which indicates the quality of kitchen with reference to sale price of the given house. If this number is high, then there is a negative relation with respect to label column that is the chance of investment failure by surprise housing company and vice-versa.

Other columns and their relationship is as follows:

In contrast, the most effective column is OverallQual that indicates the overall quality of the house with reference to its sale price which shows the positive relationship with variable target in accordance to its correlation.

**Hardware and Software Requirements and Tools Used:**

As for the hardware part the below details are as follows:

**Processor-** Intel(R) Core (TM) i5-6300 CPU @2.40GHz

**Installed Memory (RAM)-** 8.00 GB (7.82 Usable)

**System type-** 64-bit operating system, x64

**Libraries and packages used in python**:

1. Pandas
2. Numpy
3. Matplotlib.pyplot
4. Seaborn
5. sklearn.preprocessing import LabelEncoder
6. sklearn.preprocessing import StandardScaler
7. Scipy.stats import Zscore
8. sklearn.linear\_model import LinearRegression, Lasso,Ridge,ElasticNet
9. sklearn.model\_selection import train\_test\_split,cross\_val\_score,GridSearchCV
10. sklearn.neighbors import KNeighborsRegressor
11. sklearn.metrics import mean\_absolute\_error,mean\_squared\_error,r2\_score
12. sklearn.tree import DecisionTreeRegressor
13. sklearn.ensemble import RandomForestRegressor,AdaBoostRegressor,GradientBoostingRegressor,ExtraTreesRegressor,
14. sklearn.svm import SVR
15. Warnings

All these tools and libraries are used to perform the model selection process and identifying which model is best fitted to my project with giving the least errors.

**Model/s Development and Evaluation**

**Identification of possible problem-solving approaches (methods):**

In this section it includes the approach that I have used to make statistical and analytical data analysis to solve the problem statement. Firstly, I have cleaned the data, finding out the statistical observations which is based on the given dataset.

Secondly doing an exploratory data analysis which describes the behaviour of each data columns with respect to its label target column and then finding out, standard scaler in the dataset which concludes the correlation of each columns to its label target.

Lastly, detecting the outliers using Zscore method and remove those outliers to make the data out of extremities and then performing the machine learning algorithms to finding out which model is giving us the maximum accuracy with least errors as a performance metrics.

**Testing of Identified Approaches (Algorithms):**

As per this project where data is crucial to us and data size is too much, I have applied basic algorithms as it is taking so much time to process it in my computer. The list of algorithms as follows:

* Linear Regression
* Decision Tree Regressor
* KNeighbors Regressor
* Support Vector Regressor
* Lasso, Ridge, Elastic Net
* Random Forest Regressor
* Gradient boosting Regressor
* Extra trees Regressor

**Evaluation of Selected Models:**

In the evaluation phase I have concluded that our model is performing and giving the best results on RandomFrorest Regressor with the least number of errors.

In other Regression models I am getting the accuracy score of more than 70% but, I have observed that the model is getting the maximum accuracy score of more than 97.99% which is good that we have attained the accuracy score of more than 97.99% with the least counts of errors as per the r2 score,mean squared error and mean absolute error for the given dataset.

**Key Metrics for success in solving problem under consideration:**

As per the key performance metrics I have used accuracy score,R2 score, mean squared error, root mean squared error and mean absolute error to determine the best fitted model as per the given dataset.

Key observations includes the working of best model with the algorithm of Random Forest Regressor which given the maximum accuracy with the given performance metrics.

As for the statistical metrics I have used the Zscore function to detect the outliers and correlation with respect to our target label in the given dataset. Both the method have performed well in showing the required results.

**Visualizations:**

The plots and figures are as follows for the given dataset:

Chart, histogram

Description automatically generated

Mean of sale price- 181477.0059931507

This indicates that mean sale price for the given dataset is 181477.00 exactly which would be consider at the time of making investment by the housing company.

2. sale price vs sale condition Chart, box and whisker chart

Description automatically generated

This figure indication the pricing as per the sale condition of a house adjustment of land indicates the lowest sale price whereas partial house indicates the highest sale price.

3. sale price vs Garage type.

Chart, box and whisker chart

Description automatically generated

This figure indicates that Builtin garage type cost higher as compared to carport garage type.

4. sale price vs central air.

Chart, box and whisker chart

Description automatically generated

This figure indicates house in which the facility of central air is there cost higher than the ones which don’t have the central air facility.

5. sale price vs lotconfig.

Chart, box and whisker chart

Description automatically generated

This Figure indicates the lot configuration as per the sale price of the house.

1. sale price vs fireplaces.

Chart, box and whisker chart

Description automatically generated

This figure indicates sale price of the house as houses in which the fireplace is high then it cost more than the house which don’t have the fireplaces installed.

1. sale price vs house style.

Chart, box and whisker chart

Description automatically generated

This figure indicate the sale price affected by the housing style for the given dataset.

1. Checking the correlation with the target variable.

Chart, bar chart

Description automatically generated

This figure indicates the correlation between different factors which determine the sale price for the given house. OverallQual is the highest affecting factor in the figure whereas kichenAbvGr is the lowest affecting factor with the target variable.

**CONCLUSION**

**Key Findings and Conclusions of the Study:**

Houses are one of the necessary need of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors in the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

As In the given project there is a lot to know as the project was given to find whether to make a fruitful investment based upon the no. of factors that are affecting the sale price of a house.

While working in the project I come to know that most of the data is right side correlated and every other column is affecting the label target variable in huge sense expect the columns where data is negatively correlated such as in the column of KichenAbvGr, Enclosedporch etc.

The whole states the success and failure of Investment amount which will be predicted using different machine learning algorithms so that in future there is chance of implementing different techniques to manage the no. of factors which is affecting the real time data now.