

|  |
| --- |
| **HOUSE PRICE PREDICTION PROJECT** |
| **Prepared by**  **Arti Sharma**  Data Science Intern at FlipRobo Technologies    SME Name: Sapna Verma |

**HOUSE PRICE PREDICTION PROJECT pg. 1**

|  |  |  |
| --- | --- | --- |
| **Acknowledgement** | | |
|  | | |
|  | It is my deepest pleasure and gratification to present this report. Working on this project was an incredible experience that has given me a very informative knowledge regarding the data analysis process.  All the required information and dataset are provided by **Flip Robo Technologies** (Bangalore) that helped me to complete the project. I want to thank my SME **Sapna Verma** for giving the dataset and instructions to perform the complete case study process. |  |

**HOUSE PRICE PREDICTION PROJECT pg. 2**

**INTRODUCTION**

Houses are one of the necessary needs 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 someof the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company. A US-based housing company named Surprise Housing has decided to enter the Australian market. The companyuses 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.

### Business Problem Framing:

Thousands of houses are sold every day. There are some questions every buyer asks himself like: What is the actual price that this house deserves? Am I paying a fair price? In this paper,a machine learning model is proposed to predict a house price based on data related to the house (its size, the year it was built in, etc.). During the development and evaluation of our model, we will show the code used for each step followed by its output. This will facilitate the reproducibility of our work. In this study, Python programming language with a number of Python packages will be used.

### Conceptual Background of the Domain Problem:

The main objectives of this study are as follows:

* + To apply data pre-processing and preparation techniques in order to obtainclean data
  + To build machine learning models able to predict house price based onhouse features
  + To analyse and compare model’s performance in order to choose the bestmodel

## Literature Review

Machine learning is a form of artificial intelligence which compose available

**HOUSE PRICE PREDICTION PROJECT pg. 3**

computers with the efficiency to be trained without being veraciously programmed. Machine learning interest on the extensions of computer programs which is capable enough to modify when unprotected to new-fangled data. Machine learning algorithms are broadlyclassified into three divisions, namely; Supervised learning, Unsupervised learning and Reinforcement learning. Supervised learning is a learning in which we teach or train the machine using data which is well labelled that means some data is already tagged with correct answer. After that, machine is provided with new set of examples so that supervised learning algorithm analyses the training data and produces a correct outcome from labelled data. Unsupervised learning is the training of machine using information that is neither classified nor labelled and allowing the algorithm to act on that informationwithout guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data. Unlike, supervisedlearning, no teacher is provided that means no training will be given to the machine.

Therefore, machine is restricted to find the hidden structure in unlabeled data by our-self.

Reinforcement learning is an area of Machine Learning. Reinforcement. It is about takingsuitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behavior or path it should take in a specific situation. Reinforcement learning differs from the supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of training dataset, it is bound to learn from its experience. Machine learning has many applications out of which one of the applications is prediction of real estate. The real estate market is one of the most competitive in terms of pricing and same tends to be vary significantly based on lots of factor, forecasting property price is an important modules in decision making for both the buyers and investors in supporting budget allocation, finding property finding stratagems and determining suitable policies hence it becomes one of the prime fields to apply the concepts of machine learning to optimize and predict the prices with high accuracy. The study on land price trend is felt important to support the decisions in urban planning. The real estate system is an unstable stochasticprocess. Investor’s decisions are based on the market trends to reap maximum returns.

Developers are interested to know the future trends for their decision making. To accurately estimate property prices and future trends, large amount of data that influences land price is required for analysis, modelling and forecasting. The factors that affect the land price have to be studied and their impact on price has also to be modelled. An analysis of the past data is to be considered. It is inferred that establishing a simple linear mathematical relationship for these time-series data is found not viable for forecasting.

Hence it became imperative to establish a non-linear model which can well fit the data characteristic to analyse and forecast future trends. As the real estate is fast developing sector, the analysis and forecast of land prices using mathematical modelling and other scientific techniques is an immediate urgent need for decision making by all those concerned. The increase in population as well as the industrial activity is attributed to various factors, the most prominent being the recent spurt in the knowledge sector viz. Information Technology (IT) and Information technology enabled services. Demand for land started of showing an upward trend

**HOUSE PRICE PREDICTION PROJECT pg. 4**

and housing and the real estate activity started booming. All barren lands and paddy fields ceased their existence to pave way for multistore and high-rise buildings. Investments started pouring in Real estate Industry and there was no uniform pattern in the land price over the years. The need for predicting the trend in land prices was felt by all in the industry viz. the Government, the regulating bodies, lending institutions, the developers and the investors. Therefore, in this paper, we present various important features to use while predicting housing prices with good accuracy. We can use regression models, using various features to have lower Root mean Squared error. While using features in a regression model some feature engineering is required for better prediction. Often a set of features linear regression, random forest regression and decision tree regression is used for making better model fit. For these models are expected to be susceptible towards over fitting random forest regression is used to reduce it. So, it directs to the best application of regression models in addition to other techniques to optimize the result.

**Linear Regression:**

To establish baseline performance with a linear classifier, we used Linear Regression tomodel the price targets, Y, as a linear function of the data, X

(𝑋) = 𝑤0 + 𝑤1𝑥1 + … + 𝑤m𝑥m + 𝑥m

∞

= Σ (𝑤j𝑥j)

j=l:m

Advantage: A linear model can include more than one predictor as long as the predictors are additive. the best fit line is the line with minimum error from all the points, it has highefficiency but sometimes this high efficiency created.

Disadvantage: Linear Regression Is Limited to Linear Relationships. Linear Regression Only Looks at the Mean of the Dependent Variable. Linear Regression Is Sensitive to Outliers. Data Must Be Independent

#### Random Forest Regression:

The Random Forest Regression (RFR) is an ensemble algorithm that combines multiple Regression Trees (RTs). Each RT is trained using a random subset of the features, and theoutput is the average of the individual RTs. The sum of squared errors for a tree T is:

Advantages: There is no need for feature normalization. Individual decision trees can be trained in parallel. Random forests are widely used. They reduce overfitting.

Disadvantages: They’re not easily interpretable. They’re not a state-of-the-art

𝑆 = Σ ∑(𝑦i − 𝑚c)2

c€leaves(T)i€C

Where = 1 + Σ 𝑦

c nc

i

i€C

**HOUSE PRICE PREDICTION PROJECT pg. 5**

## Analytical Problem Framing

### Mathematical/ Analytical Modelling of the Problem StatisticalAnalysis

Once it comes time to analyze the data, there are an array of statistical model’s analystsmay choose to utilize. The most common techniques will fall into the following two groups:

* Supervised learning, including regression and classification models.
* Unsupervised learning, including clustering algorithms and association rules

#### Regression Model:

The regression models are used to examine relationships between variables. Regression models are often used to determine which independent variables hold the most influence overdependent variables information that can be leveraged to make essential decision.

The most traditional regression model is linear regression, decision tree regression, randomforest regression, xgboost regression and knn-neighbours.

There are 4 main components of an analytics model, namely: 1) Data Component, 2) Algorithm Component, 3) Real World Component, and 4) Ethical Component.

## Data Preparation

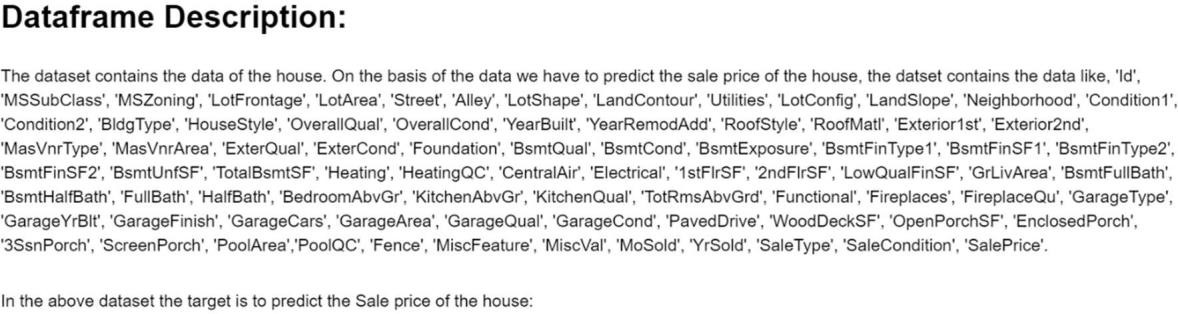
In this study, we will use a housing dataset presented by A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses dataanalytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses inAustralia. The data is provided in the csv file below housing\_train.csv and housing\_test.csv

## Data Description

The dataset contains 1460 records (rows) and 81 features (columns).

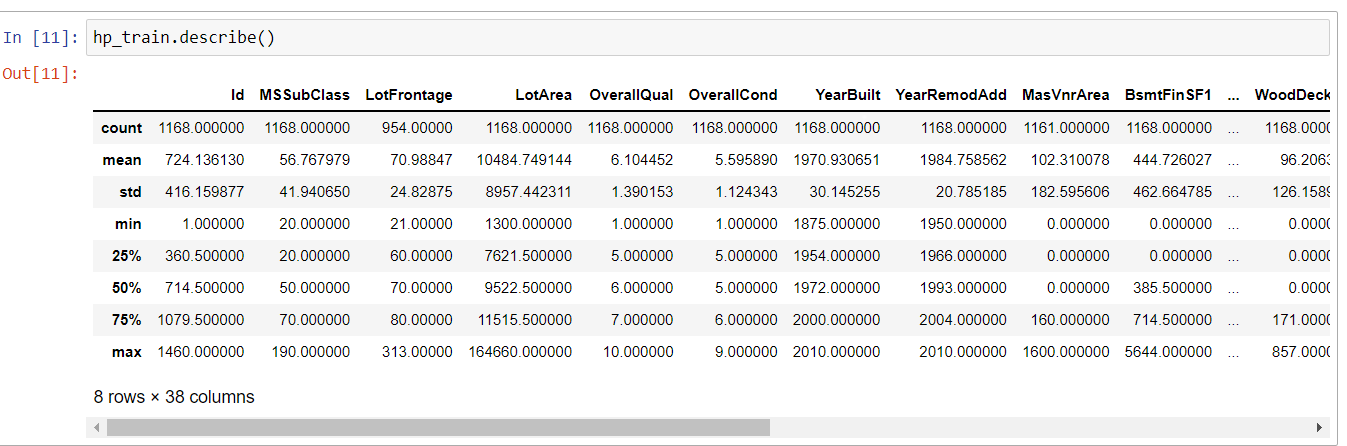
Here, we will provide a brief description of dataset features. Since the number of features islarge (81), we will attach the original data description file to this study for more informationabout the dataset. Now, we will mention the feature name with a short description of its meaning.

**HOUSE PRICE PREDICTION PROJECT pg. 6**



# Data Pre-processing

Then we move to see statistical information about the non-numerical columns in ourdataset:



From the table above, we can see, for example, that the average lot area of the houses in our dataset is 10,484.74 ft2 with a standard deviation of 8957.44 ft2. We can see also thatthe minimum lot area is 1300 ft2 and the maximum lot area is 164660 ft2 with a median of 9522 ft2. Similarly, we can get a lot of information about our dataset variables from the table.

## Data Cleaning

### Dealing with Missing Values:

We should deal with the problem of missing values because some machine learning modelsdon't accept data with missing values. Firstly, let's see the number of missing values in our dataset. We want to see the number and the percentage of missing values for each column that actually contains missing values.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Columns** | **non-null** | **Columns** | **non-null** | **Columns** | **non-null** |
| Id | 0 | ExterCond | 0 | Functional | 0 |
| MSSubClass | 0 | Foundation | 0 | Fireplaces | 0 |
| MSZoning | 0 | BsmtQual | 30 | FireplaceQu | 551 |
| LotFrontage | 214 | BsmtCond | 30 | GarageType | 64 |
| LotArea | 0 | BsmtExposure | 31 | GarageYrBlt | 64 |
| Street | 0 | BsmtFinType1 | 30 | GarageFinish | 64 |
| Alley | 1091 | BsmtFinSF1 | 0 | GarageCars | 0 |
| LotShape | 0 | BsmtFinType2 | 31 | GarageArea | 0 |
| LandContour | 0 | BsmtFinSF2 | 0 | GarageQual | 64 |
| LotConfig | 0 | BsmtUnfSF | 0 | GarageCond | 64 |
| LandSlope | 0 | TotalBsmtSF | 0 | PavedDrive | 0 |

**HOUSE PRICE PREDICTION PROJECT pg. 7**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Neighborhood | 0 | Heating | 0 | WoodDeckSF | 0 |
| Condition1 | 0 | HeatingQC | 0 | OpenPorchSF | 0 |
| Condition2 | 0 | CentralAir | 0 | EnclosedPorch | 0 |
| BldgType | 0 | Electrical | 0 | 3SsnPorch | 0 |
| HouseStyle | 0 | 1stFlrSF | 0 | ScreenPorch | 0 |
| OverallQual | 0 | 2ndFlrSF | 0 | PoolArea | 0 |
| OverallCond | 0 | LowQualFinSF | 0 | PoolQC | 1161 |
| YearBuilt | 0 | GrLivArea | 0 | Fence | 931 |
| YearRemodAdd | 0 | BsmtFullBath | 0 | MiscFeature | 1124 |
| RoofStyle | 0 | BsmtHalfBath | 0 | MiscVal | 0 |
| RoofMatl | 0 | FullBath | 0 | MoSold | 0 |
| Exterior1st | 0 | HalfBath | 0 | YrSold | 0 |
| Exterior2nd | 0 | BedroomAbvGr | 0 | SaleType | 0 |
| MasVnrType | 7 | KitchenAbvGr | 0 | SaleCondition | 0 |
| MasVnrArea | 7 | KitchenQual | 0 | SalePrice | 0 |
| ExterQual | 0 | TotRmsAbvGrd | 0 |  | |

In the above table we got, count represents the number of non-null values in each column.Filling the missing values using fillna method.



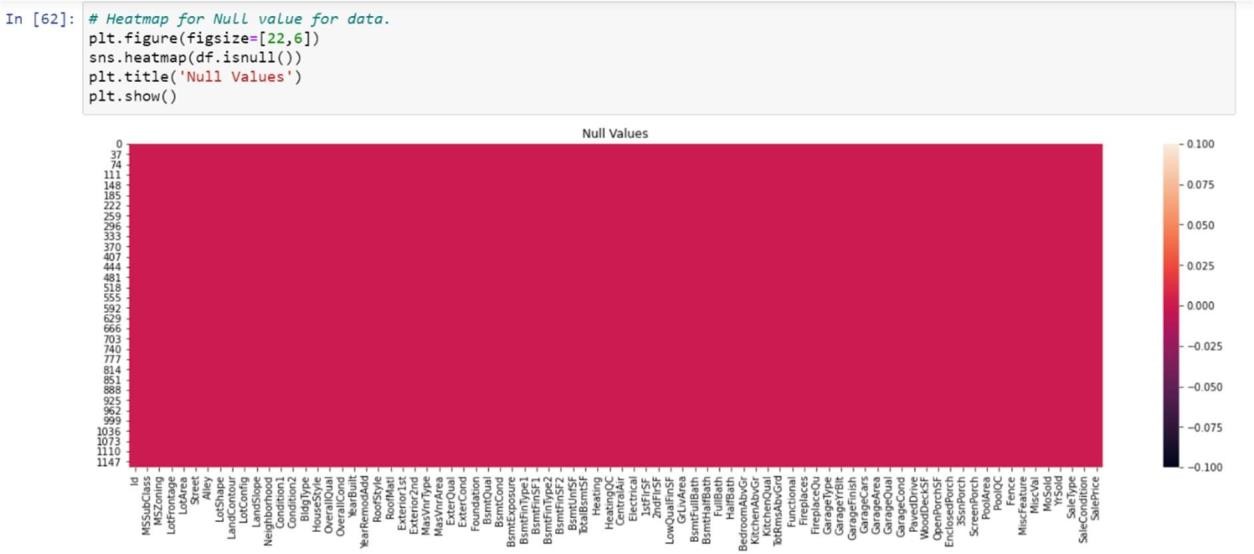
Now let's check if there is any remaining missing value in our dataset:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Columns** | **non**  **-**  **nul**  **l** | **Columns** | **non- null** | **Columns** | **non- null** |
| Id | 0 | ExterCond | 0 | Functional | 0 |
| MSSubClass | 0 | Foundation | 0 | Fireplaces | 0 |
| MSZoning | 0 | BsmtQual | 0 | FireplaceQu | 0 |
| LotFrontage | 0 | BsmtCond | 0 | GarageType | 0 |
| LotArea | 0 | BsmtExposure | 0 | GarageYrBlt | 0 |
| Street | 0 | BsmtFinType1 | 0 | GarageFinish | 0 |
| Alley | 0 | BsmtFinSF1 | 0 | GarageCars | 0 |
| LotShape | 0 | BsmtFinType2 | 0 | GarageArea | 0 |

**HOUSE PRICE PREDICTION PROJECT pg. 8**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| LandContour | 0 | BsmtFinSF2 | 0 | GarageQual | 0 |
| LotConfig | 0 | BsmtUnfSF | 0 | GarageCond | 0 |
| LandSlope | 0 | TotalBsmtSF | 0 | PavedDrive | 0 |
| Neighborhood | 0 | Heating | 0 | WoodDeckSF | 0 |
| Condition1 | 0 | HeatingQC | 0 | OpenPorchSF | 0 |
| Condition2 | 0 | CentralAir | 0 | EnclosedPorch | 0 |
| BldgType | 0 | Electrical | 0 | 3SsnPorch | 0 |
| HouseStyle | 0 | 1stFlrSF | 0 | ScreenPorch | 0 |
| OverallQual | 0 | 2ndFlrSF | 0 | PoolArea | 0 |
| OverallCond | 0 | LowQualFinSF | 0 | PoolQC | 0 |
| YearBuilt | 0 | GrLivArea | 0 | Fence | 0 |
| YearRemodAdd | 0 | BsmtFullBath | 0 | MiscFeature | 0 |
| RoofStyle | 0 | BsmtHalfBath | 0 | MiscVal | 0 |
| RoofMatl | 0 | FullBath | 0 | MoSold | 0 |
| Exterior1st | 0 | HalfBath | 0 | YrSold | 0 |
| Exterior2nd | 0 | BedroomAbvGr | 0 | SaleType | 0 |
| MasVnrType | 0 | KitchenAbvGr | 0 | SaleCondition | 0 |
| MasVnrArea | 0 | KitchenQual | 0 | SalePrice | 0 |
| ExterQual | 0 | TotRmsAbvGrd | 0 |  | |

This means that our dataset is now complete; it doesn't contain any missing value anymore.To show graphical representation of null using heatmap for entire dataset.



|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Data types** | **Column** | **Data types** |
| Id | int64 | CentralAir | object |
| MSSubClass | int64 | Electrical | object |
| MSZoning | object | 1stFlrSF | int64 |
| LotFrontage | float64 | 2ndFlrSF | int64 |
| LotArea | int64 | LowQualFinSF | int64 |
| Street | object | GrLivArea | int64 |
| Alley | object | BsmtFullBath | int64 |
| LotShape | object | BsmtHalfBath | int64 |
| LandContour | object | FullBath | int64 |
| LotConfig | object | HalfBath | int64 |
| LandSlope | object | BedroomAbvGr | int64 |
| Neighborhood | object | KitchenAbvGr | int64 |
| Condition1 | object | KitchenQual | object |
| Condition2 | object | TotRmsAbvGrd | int64 |
| BldgType | object | Functional | object |

**HOUSE PRICE PREDICTION PROJECT pg. 9**

|  |  |  |  |
| --- | --- | --- | --- |
| HouseStyle | object | Fireplaces | int64 |
| OverallQual | int64 | FireplaceQu | object |
| OverallCond | int64 | GarageType | object |
| YearBuilt | int64 | GarageYrBlt | float64 |
| YearRemodAdd | int64 | GarageFinish | object |
| RoofStyle | object | GarageCars | int64 |
| RoofMatl | object | GarageArea | int64 |
| Exterior1st | object | GarageQual | object |
| Exterior2nd | object | GarageCond | object |
| MasVnrType | object | PavedDrive | object |
| MasVnrArea | float64 | WoodDeckSF | int64 |
| ExterQual | object | OpenPorchSF | int64 |
| ExterCond | object | EnclosedPorch | int64 |
| Foundation | object | 3SsnPorch | int64 |
| BsmtQual | object | ScreenPorch | int64 |
| BsmtCond | object | PoolArea | int64 |
| BsmtExposure | object | PoolQC | object |
| BsmtFinType1 | object | Fence | object |
| BsmtFinSF1 | int64 | MiscFeature | object |
| BsmtFinType2 | object | MiscVal | int64 |
| BsmtFinSF2 | int64 | MoSold | int64 |
| BsmtUnfSF | int64 | YrSold | int64 |
| TotalBsmtSF | int64 | SaleType | object |
| Heating | object | SaleCondition | object |
| HeatingQC | object | SalePrice | int64 |

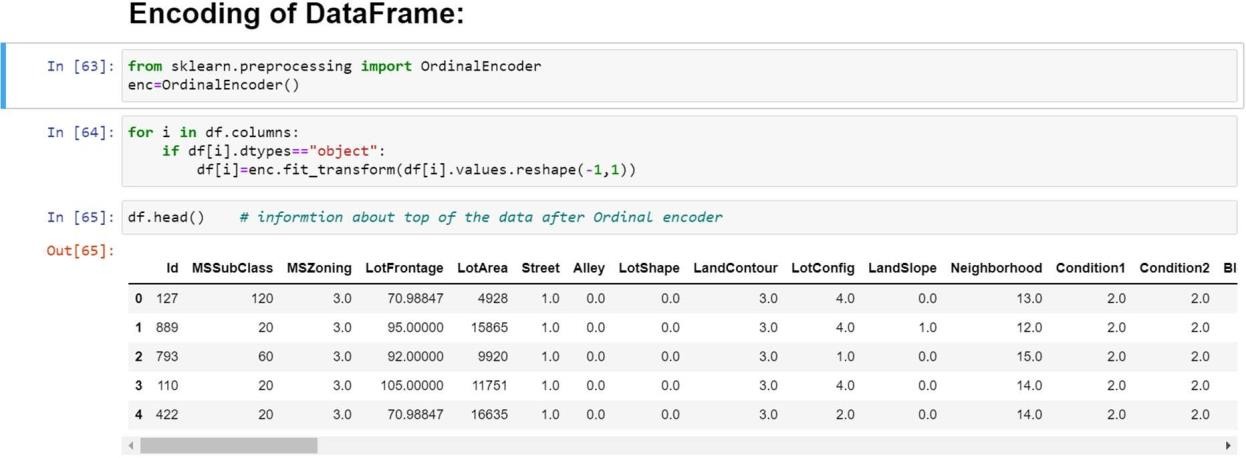
Since the dataset has a lot string values. We will use the encoding techniques to convert thestring data to numerical one.

### Encoding of Data Frame:

In ordinal encoding, each unique category value is assigned an integer value. This ordinal encoding transform is available in the scikit-learn Python machine learning library via the Ordinal Encoder class. By default, it will assign integers to labels in the order that is observedin the data.

In this encoding scheme, the categorical feature is first converted into numerical using an ordinal encoder. Then the numbers are transformed in the binary number. After that binary value is split into different columns. Binary encoding works really well when there are a highnumber of categories.

**HOUSE PRICE PREDICTION PROJECT pg. 10**



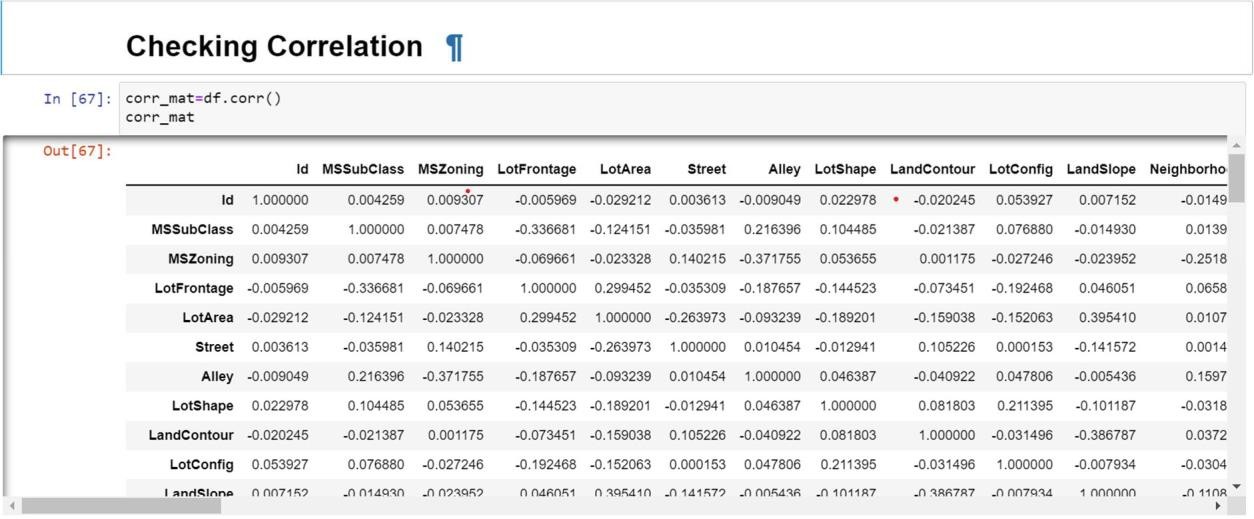
* **Correlation matrix:**

A correlation matrix is simply a table which displays the correlation. The measure is best used in variables that demonstrate a linear relationship between each other. The fit of the datacan be visually represented in a heatmap.

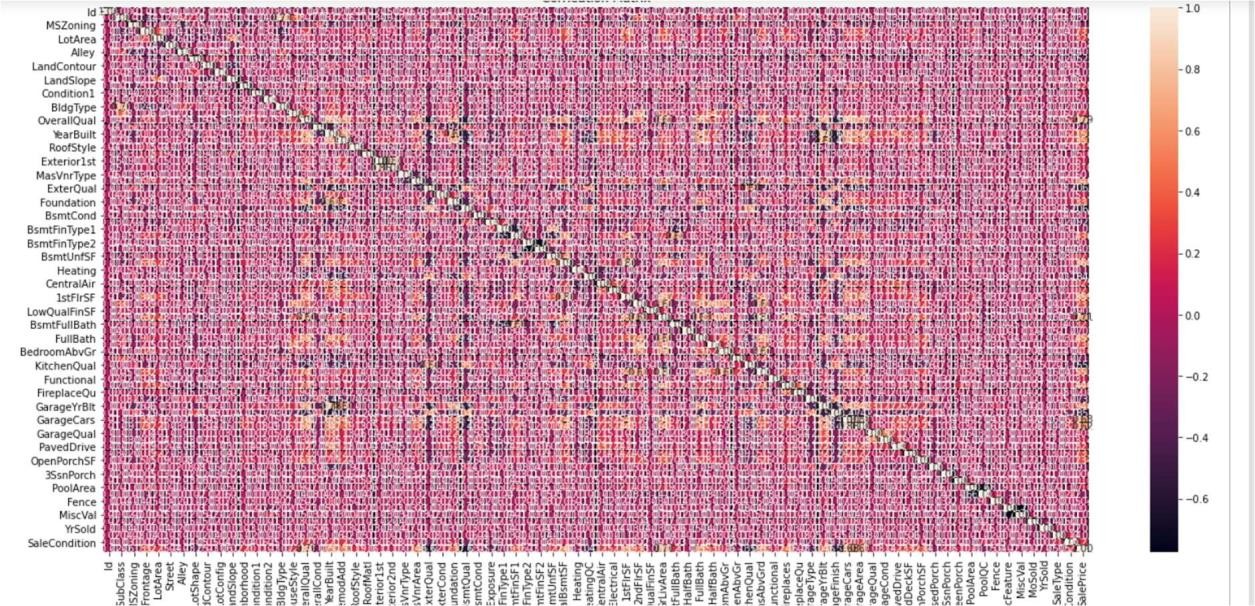
Pandas dataframe. corr() method is used for creating the correlation matrix. It is used to findthe pairwise correlation of all columns in the dataframe.

To create correlation matrix using pandas, these steps should be taken:

1. Obtain the data.
2. Create the DataFrame using Pandas.
3. Create correaltion matrix using Pandas.



**HOUSE PRICE PREDICTION PROJECT pg. 11**



Observations: We are unable to identify the correlation in above heatmap due to huge numberof columns.

How correlation matrix is calculated?

A correlation matrix is a table showing correlation coefficients between sets of variables. Each random variable (x) in the table is correlated with each of the other values in the table x.The diagonal of the table is always a set of ones, because the correlation between a variableand itself is always 1.



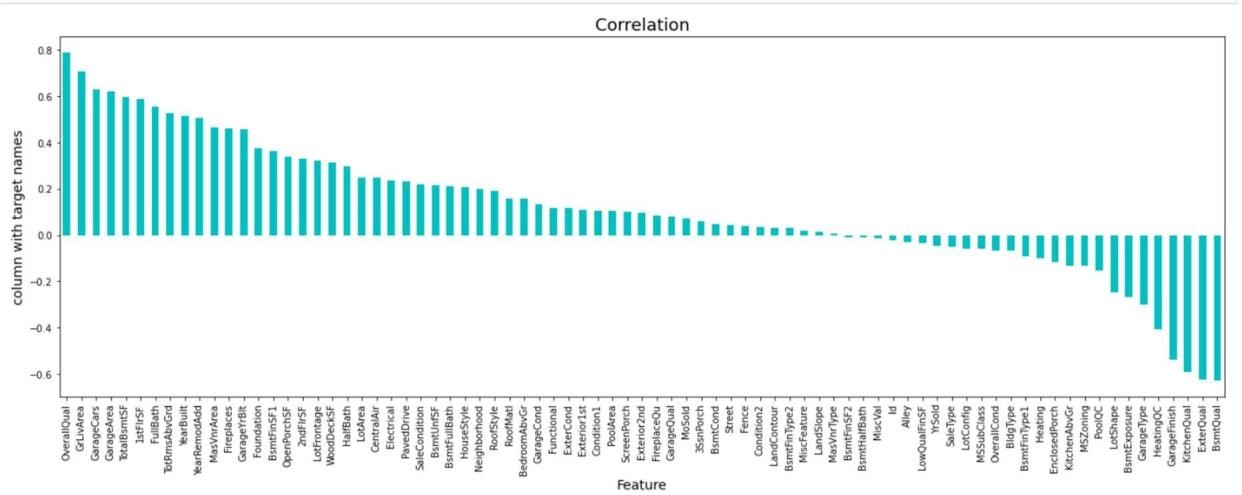
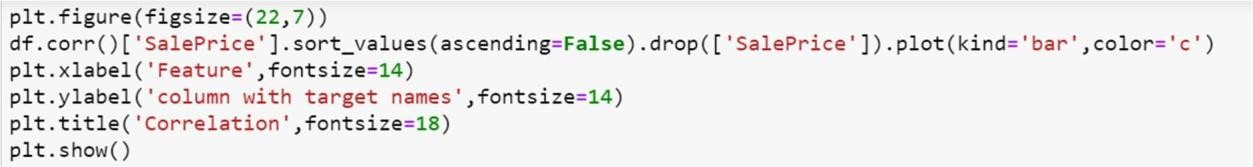
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Columns** | **Correlation matrix** | **Columns** | **Correlation** | **matrix** |
| SalePrice | 1.000000 | ExterCond | 0.115167 |  |
| OverallQual | 0.789185 | Exterior1st | 0.108451 |  |
| GrLivArea | 0.707300 | Condition1 | 0.105820 |  |
| GarageCars | 0.628329 | PoolArea | 0.103280 |  |
| GarageArea | 0.619000 | ScreenPorch | 0.100284 |  |
| TotalBsmtSF | 0.595042 | Exterior2nd | 0.097541 |  |
| 1stFlrSF | 0.587642 | LandSlope | 0.015485 |  |
| FullBath | 0.554988 | MasVnrType | 0.007732 |  |
| TotRmsAbvGrd | 0.528363 | BsmtFinSF2 | -0.010151 |  |
| YearBuilt | 0.514408 | BsmtHalfBath | -0.011109 |  |
| YearRemodAdd | 0.507831 | MiscVal | -0.013071 |  |
| MasVnrArea | 0.463626 | Id | -0.023897 |  |
| Fireplaces | 0.459611 | Alley | -0.029798 |  |
| GarageYrBlt | 0.458007 | LowQualFinSF | -0.032381 |  |
| Foundation | 0.374169 | YrSold | -0.045508 |  |
| BsmtFinSF1 | 0.362874 | SaleType | -0.050851 |  |
| OpenPorchSF | 0.339500 | LotConfig | -0.060452 |  |
| 2ndFlrSF | 0.330386 | MSSubClass | -0.060775 |  |
| LotFrontage | 0.323779 | OverallCond | -0.065642 |  |
| WoodDeckSF | 0.315444 | BldgType | -0.066028 |  |
| HalfBath | 0.295592 | BsmtFinType1 | -0.092109 |  |
| LotArea | 0.249499 | Heating | -0.100021 |  |
| CentralAir | 0.246754 | EnclosedPorch | -0.115004 |  |
| Electrical | 0.234621 | KitchenAbvGr | -0.132108 |  |

**HOUSE PRICE PREDICTION PROJECT pg. 12**

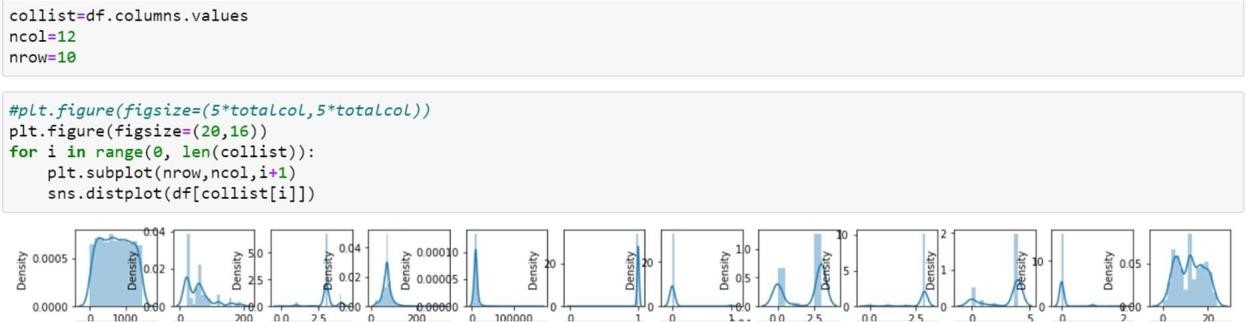
|  |  |  |  |
| --- | --- | --- | --- |
| PavedDrive | 0.231707 | MSZoning | -0.133221 |
| SaleCondition | 0.217687 | PoolQC | -0.152611 |
| BsmtUnfSF | 0.215724 | LotShape | -0.248171 |
| BsmtFullBath | 0.212924 | BsmtExposure | -0.268559 |
| HouseStyle | 0.205502 | GarageType | -0.299470 |
| Neighborhood | 0.198942 | HeatingQC | -0.406604 |
| RoofStyle | 0.192654 | GarageFinish | -0.537121 |
| RoofMatl | 0.159865 | KitchenQual | -0.592468 |
| BedroomAbvGr | 0.158281 | ExterQual | -0.624820 |
| GarageCond | 0.135071 | BsmtQual | -0.626850 |
| Functional | 0.118673 |  | |

Now we can clearly identify the correlation of independent variables with the target variables "Sale Price". There are around 25 variables who has less than 0.01 correlation value (very week relationship.)

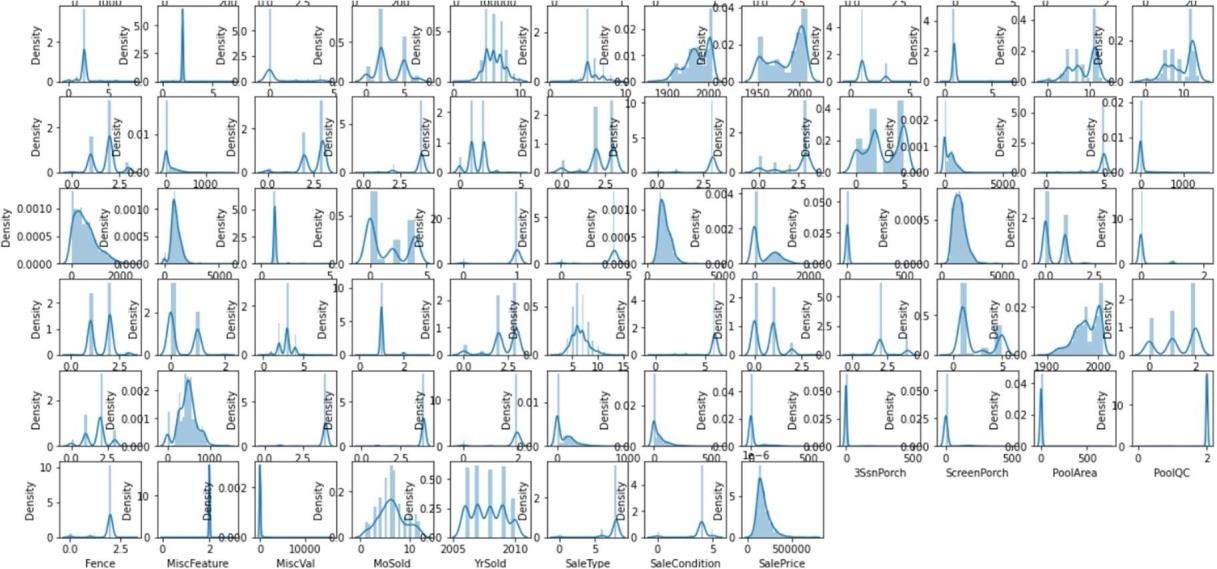
Checking the columns which are positively and negative correlated with the target columns:



Let’s check the data distribution among all the columns.



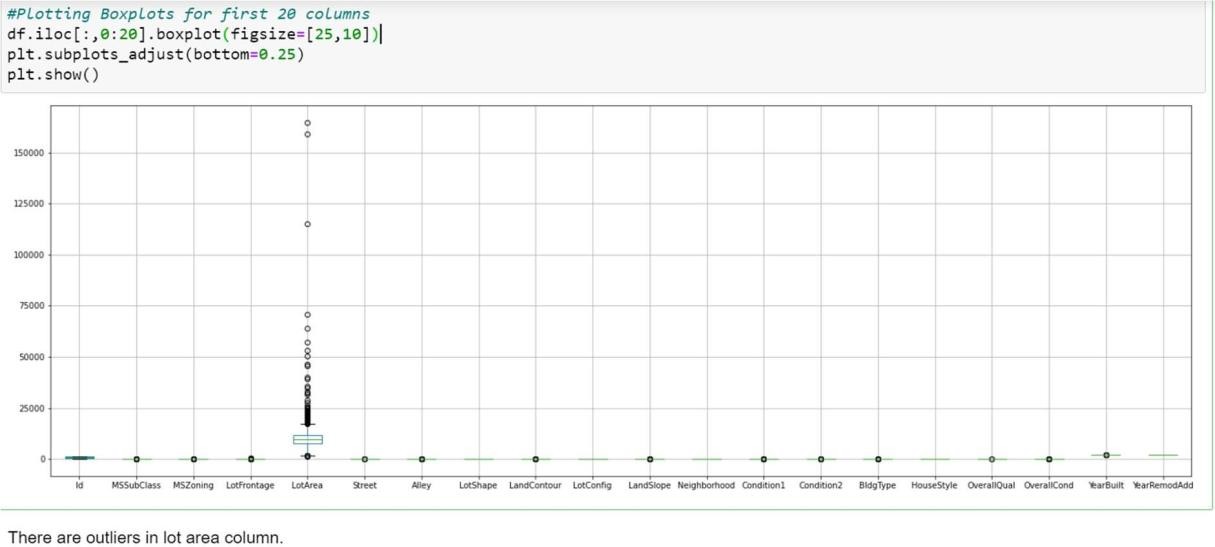
**HOUSE PRICE PREDICTION PROJECT pg. 13**



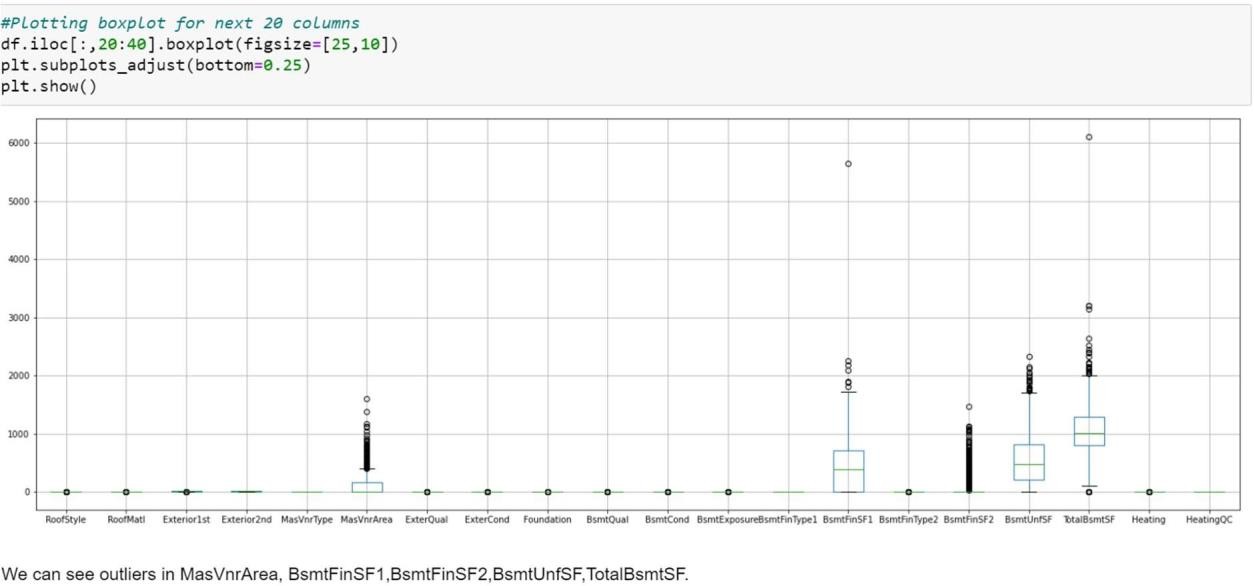
We can see skewness in data for the multiple columns, will handle the skewness in furthersteps.

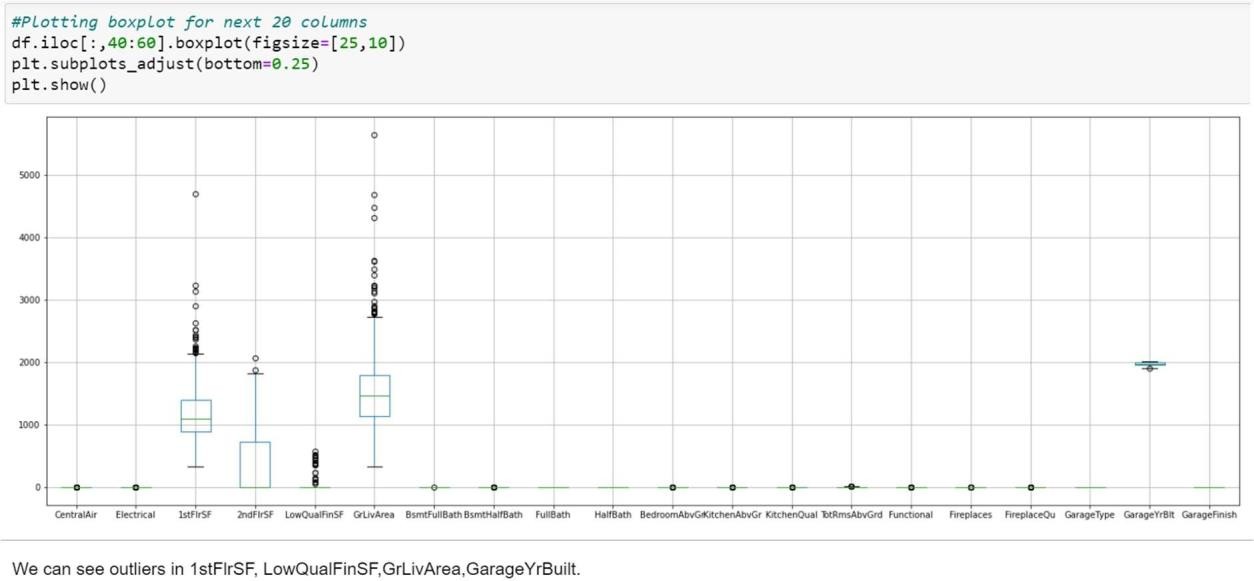
## Outliers Check:

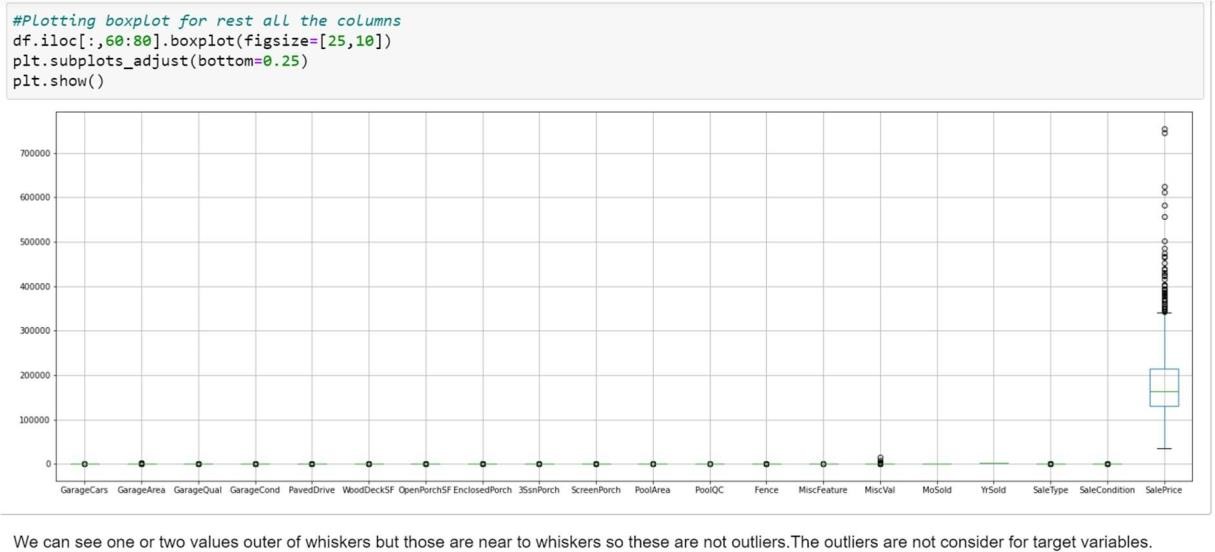
There are 80 columns in dataset so it’s not possible to plot each and every column separatelyor plot all together. so, we will print in 4 steps:



**HOUSE PRICE PREDICTION PROJECT pg. 14**







**HOUSE PRICE PREDICTION PROJECT pg. 15**

* **Skewness:**

Skewness is a measure of symmetry in a distribution. Actually, it's more correct to describe it as a measure of lack of symmetry. A standard normal distribution is perfectly symmetrical and has zero skew. Therefore, we need a way to calculate how much the distribution is skewed.

* + Checking Skewness:

|  |  |  |  |
| --- | --- | --- | --- |
| **Columns** | **Skewness** | **Columns** | **Skewness** |
| Id | 0.026526 | CentralAir | -3.475188 |
| MSSubClass | 1.422019 | Electrical | -3.104209 |
| MSZoning | -1.796785 | 1stFlrSF | 1.513707 |
| LotFrontage | 2.710383 | 2ndFlrSF | 0.823479 |
| LotArea | 10.659285 | LowQualFinSF | 8.666142 |
| Street | -17.021969 | GrLivArea | 1.449952 |
| Alley | 5.436187 | BsmtFullBath | 0.627106 |
| LotShape | -0.603775 | BsmtHalfBath | 4.264403 |
| LandContour | -3.125982 | FullBath | 0.057809 |
| LotConfig | -1.118821 | HalfBath | 0.656492 |
| LandSlope | 4.812568 | BedroomAbvGr | 0.243855 |
| Neighborhood | 0.043735 | KitchenAbvGr | 4.365259 |
| Condition1 | 3.008289 | KitchenQual | -1.408106 |
| Condition2 | 11.514458 | TotRmsAbvGrd | 0.644657 |
| BldgType | 2.318657 | Functional | -3.999663 |
| HouseStyle | 0.285680 | Fireplaces | 0.671966 |
| OverallQual | 0.175082 | FireplaceQu | 0.753507 |
| OverallCond | 0.580714 | GarageType | 0.831142 |
| YearBuilt | -0.579204 | GarageYrBlt | -0.662934 |
| YearRemodAdd | -0.495864 | GarageFinish | -0.450190 |
| RoofStyle | 1.498560 | GarageCars | -0.358556 |
| RoofMatl | 7.577352 | GarageArea | 0.189665 |
| Exterior1st | -0.612816 | GarageQual | -4.582386 |
| Exterior2nd | -0.592349 | GarageCond | -5.422472 |
| MasVnrType | -0.104609 | PavedDrive | -3.274035 |
| MasVnrArea | 2.834658 | WoodDeckSF | 1.504929 |
| ExterQual | -1.810843 | OpenPorchSF | 2.410840 |
| ExterCond | -2.516219 | EnclosedPorch | 3.043610 |
| Foundation | -0.002761 | 3SsnPorch | 9.770611 |
| BsmtQual | -1.343781 | ScreenPorch | 4.105741 |
| BsmtCond | -3.293554 | PoolArea | 13.243711 |
| BsmtExposure | -1.166987 | PoolQC | -19.401558 |
| BsmtFinType1 | -0.068901 | Fence | -3.185107 |
| BsmtFinSF1 | 1.871606 | MiscFeature | -17.238424 |
| BsmtFinType2 | -3.615783 | MiscVal | 23.065943 |
| BsmtFinSF2 | 4.365829 | MoSold | 0.220979 |
| BsmtUnfSF | 0.909057 | YrSold | 0.115765 |
| TotalBsmtSF | 1.744591 | SaleType | -3.660513 |
| Heating | 10.103609 | SaleCondition | -2.671829 |
| HeatingQC | 0.449933 | SalePrice | 1.953878 |

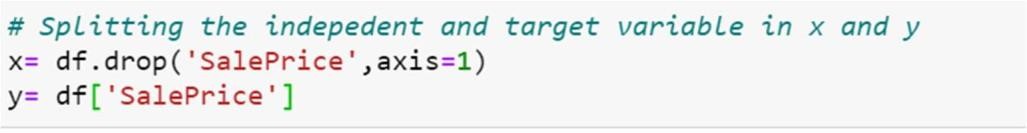
To handle skewness of the data using different types of functions:

1. Log Transform

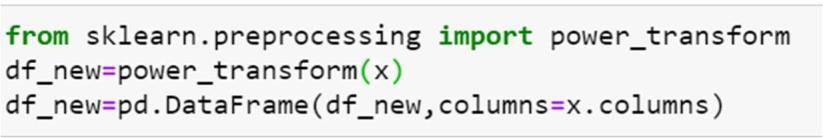
**HOUSE PRICE PREDICTION PROJECT pg. 16**

1. Square Root Transform
2. Box-Cox Transform
3. Power transform

Now here, we are going to use Power transform function to handle skewness in dataset.Then, splitting the independent and target variable in x and y.



In statistics, a power transform is a family of functions applied to create a monotonic transformation of data using power functions. It is a data transformation technique used tostabilize variance, make the data more normal distribution-like, improve the validity of measures of association (such as the Pearson correlation between variables), and for other data stabilization procedures.



After performing such statistics, the skewness is removed in dataset as shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Columns** | **Skewness** | **Columns** | **Skewness** |
| Id | -0.268486 | CentralAir | -3.475188 |
| MSSubClass | 0.064007 | Electrical | -3.006845 |
| MSZoning | 0.233113 | 1stFlrSF | -0.002391 |
| LotFrontage | 0.161368 | 2ndFlrSF | 0.280208 |
| LotArea | 0.032509 | LowQualFinSF | 6.922843 |
| Street | -17.021969 | GrLivArea | -0.000054 |
| Alley | 5.436187 | BsmtFullBath | 0.365488 |
| LotShape | -0.594207 | BsmtHalfBath | 3.954345 |
| LandContour | -2.592303 | FullBath | -0.045944 |
| LotConfig | -1.030401 | HalfBath | 0.498003 |
| LandSlope | 3.954345 | BedroomAbvGr | 0.116498 |
| Neighborhood | -0.146541 | KitchenAbvGr | -2.370593 |
| Condition1 | 0.225468 | KitchenQual | -0.435558 |
| Condition2 | 0.537277 | TotRmsAbvGrd | 0.002332 |
| BldgType | 1.857194 | Functional | -3.343664 |
| HouseStyle | -0.080331 | Fireplaces | 0.084950 |
| OverallQual | 0.021658 | FireplaceQu | 0.082653 |
| OverallCond | 0.048063 | GarageType | 0.222501 |
| YearBuilt | -0.126641 | GarageYrBlt | -0.132523 |
| YearRemodAdd | -0.225131 | GarageFinish | -0.335248 |
| RoofStyle | -0.292233 | GarageCars | -0.022970 |
| RoofMatl | -6.314987 | GarageArea | -0.320370 |
| Exterior1st | -0.338023 | GarageQual | -4.327379 |
| Exterior2nd | -0.352793 | GarageCond | -4.925781 |
| MasVnrType | -0.016203 | PavedDrive | -3.025809 |
| MasVnrArea | 0.416370 | WoodDeckSF | 0.113026 |

**HOUSE PRICE PREDICTION PROJECT pg. 17**

|  |  |  |  |
| --- | --- | --- | --- |
| ExterQual | -0.605112 | OpenPorchSF | -0.002749 |
| ExterCond | -2.270791 | EnclosedPorch | 2.022616 |
| Foundation | 0.004296 | 3SsnPorch | 7.087955 |
| BsmtQual | -0.413999 | ScreenPorch | 3.067153 |
| BsmtCond | -3.025865 | PoolArea | 12.817372 |
| BsmtExposure | -0.914214 | PoolQC | -17.021969 |
| BsmtFinType1 | -0.206639 | Fence | 1.116688 |
| BsmtFinSF1 | -0.404528 | MiscFeature | 9.291637 |
| BsmtFinType2 | -2.420885 | MiscVal | 4.991071 |
| BsmtFinSF2 | 2.394737 | MoSold | -0.035838 |
| BsmtUnfSF | -0.284390 | YrSold | 0.112893 |
| TotalBsmtSF | 0.286779 | SaleType | -2.067563 |
| Heating | -4.541694 | SaleCondition | -0.353292 |
| HeatingQC | 0.156511 |  | |

## Hardware and Software Requirements and Tools Used

### PYTHON Jupyter Notebook:

#### Key Features:

An open-source solution that has simple coding processes and syntax so it’s fairly easy tolearn Integration with other languages such as C/C++, Java, PHP, C#, etc.

Advanced analysis processes through machine learning and text mining.

Python is extremely accessible to code in comparison to other popular languages such as Java, and its syntax is relatively easy to learn making this tool popular among users that look for an open-source solution and simple coding processes. In data analysis, Python is used for data crawling, cleaning, modelling, and constructing analysis algorithms based on business scenarios. One of the best features is actually its user- friendliness: programmers don’t need toremember the architecture of the system nor handle the memory – Python is considered a high-level language that is not subject to the computer’s local processor.

#### Libraries and Packages used:

**Matplotlib:**

Matplotlib is a Python library that uses Python Script to write 2-dimensional graphs and plots. Often mathematical or scientific applications require more than single axes in a representation. This library helps us to build multiple plots at a time. You can, however, useMatplotlib to manipulate different characteristics of figures as well.

The task carried out is visualization of dataset i.e., nominal data, ordinal data, continuous data, heatmap display distribution for correlation matrix and null values, boxplot distribution for checking outliers, scatter plot distribution for modelling approach, subplot distribution foranalysis and comparison, feature importance and common importance features, line plot for prediction values vs actual values.

**HOUSE PRICE PREDICTION PROJECT pg. 18**

#### Numpy:

Numpy is a popular array – processing package of Python. It provides good support for different dimensional array objects as well as for matrices. Numpy is not only confined to providing arrays only, but it also provides a variety of tools to manage these arrays. It is fast,efficient, and really good for managing matrices and arrays.

The Numpy is used to managing matrices i.e., MAE, MSE and RMSE and arrays i.e., described the values of train test dataset.

#### Pandas:

Pandas is a python software package. It is a must to learn for data-science and dedicatedly written for Python language. It is a fast, demonstrative, and adjustable platform that offers intuitive data-structures. You can easily manipulate any type of data such as – structured ortime-series data with this amazing package.

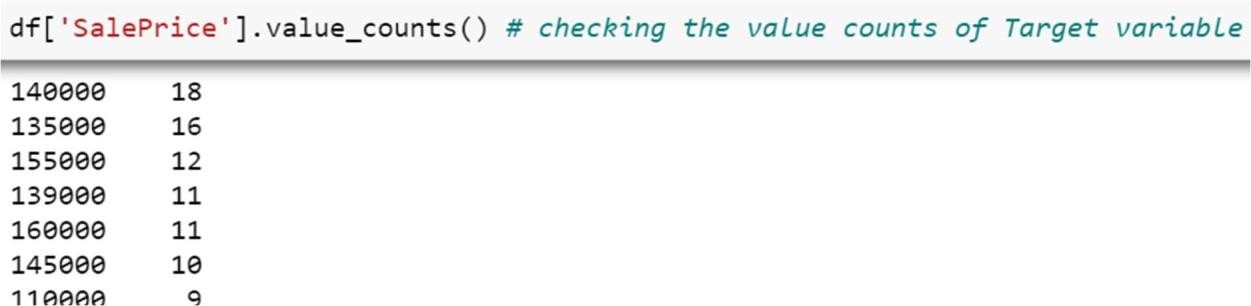
The Pandas is used to execute a Data frame i.e., test set.csv, train set.csv, skewness, co-efficient, predicted values of model approach, conclusion.

#### Scikit Learn:

Scikit learn is a simple and useful python machine learning library. It is written in python, cython, C, and C++. However, most of it is written in the Python programming language. It isa free machine learning library. It is a flexible python package that can work in complete harmony with other python libraries and packages such as Numpy and Scipy.

Scikit learn library is used to import a pre-processing function i.e., power transform, ordinal encoder, minimax scaler, linear, random forest, decision tree, xgboost, k- nearest neighbours,r2 score, mean absolute error, mean squared error, train test split, grid search cv and ensemble technique.

## Models Development and Evaluation

In this section, we choose the type of machine learning prediction that is suitable to our problem. We want to determine if this is a regression problem or a classification problem. Inthis project, we want to predict the price of a house given information about it. The price wewant to predict is a continuous value; it can be any real number. This can be seen by lookingat the target variable in our dataset Sale Price:

That means that the prediction type that is appropriate to our problem is regression.

**HOUSE PRICE PREDICTION PROJECT pg. 19**

Now, we move to choose the modelling techniques we want to use. There are a lot of techniques

**HOUSE PRICE PREDICTION PROJECT pg. 20**

available for regression problems like Linear Regression, Decision Trees, Random Forest, XGBoost, k-nearest neighbors (KNN) etc. In this project, we will test many modelling techniques, and then choose the technique(s) that yield the best results. The techniques thatwe will try are:

1. **Linear Regression**

This technique models the relationship between the target variable and the independent variables (predictors). It fits a linear model with coefficients to the data in order to minimizethe residual sum of squares between the target variable in the dataset, and the predicted values by the linear approximation.

#### Random Forest

Bagging is an ensemble method where many base models are used with a randomized subsetof data to reduce the variance of the base model.

#### Decision Trees

For this technique, the goal is to create a model that predicts the value of a target variable bylearning simple decision rules inferred from the data features.

Each one of these techniques has many algorithmic implementations. We will choosealgorithm(s) for each of these techniques in the next section.

#### k-nearest neighbors (KNN)

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine

learning algorithm that can be used to solve both classification and regression problems.

1. **AdaBoost**

AdaBoost is an ensemble learning method (also known as “meta-learning”) which was initially created to increase the efficiency of binary classifiers.

1. **GradientBoosting**

Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error.

## Model Building and Evaluation

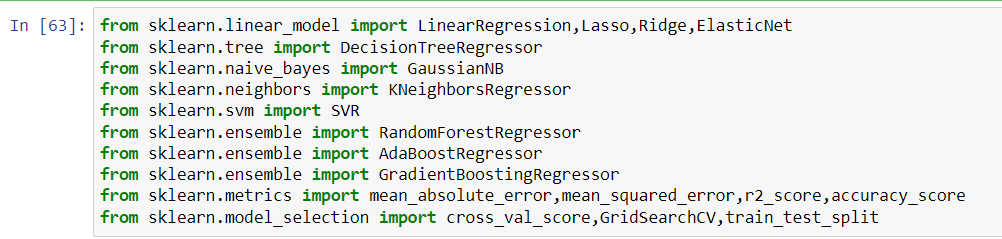
In this part, we will build our prediction model: we will choose algorithms for each of the techniques we mentioned in the previous section. After we build the model, we will evaluateits performance and results.

### Feature Scaling:

In order to make all algorithms work properly with our data, A way to normalize the input features/variables is the Min-Max scaler. By doing so, all features will be transformed into the range [0,1] meaning that the minimum and maximum value of a feature/variable is going to be 0 and 1, respectively.

**HOUSE PRICE PREDICTION PROJECT pg. 21**

**Importing libraries for metrics and model building:**

****

### Modelling Approach:

For each one of the techniques mentioned in the previous section (Linear Regression ,Random Forest Regression, Decision Tree Regression, AdaBoost, k-nearest neighbors(KNN),GradientBoosting etc.), we will follow these steps to build a model:

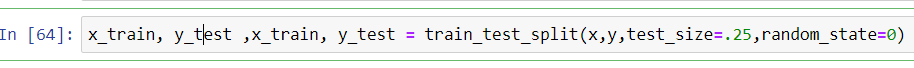
* Choose an algorithm that implements the corresponding technique
* Search for an effective parameter combination for the chosen algorithm
* Create a model using the found parameters
* Train (fit) the model on the training dataset
* Test the model on the test dataset and get the results

#### Regression Method:

* + Using Scikit-Learn, we can build a model for Linear Regression Model

### Splitting the Dataset:

As usual for supervised machine learning problems, we need a training dataset to train our model and a test dataset to evaluate the model. So, we will split our data set randomly into two parts, one for training and the other for testing. For that, we will use another function from Scikit-Learn called train\_test\_split():

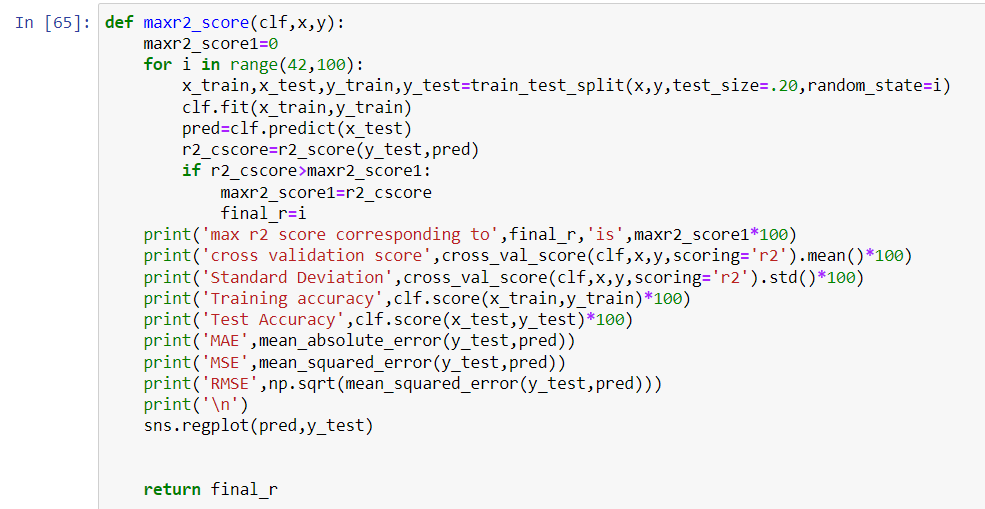


**HOUSE PRICE PREDICTION PROJECT pg. 22**

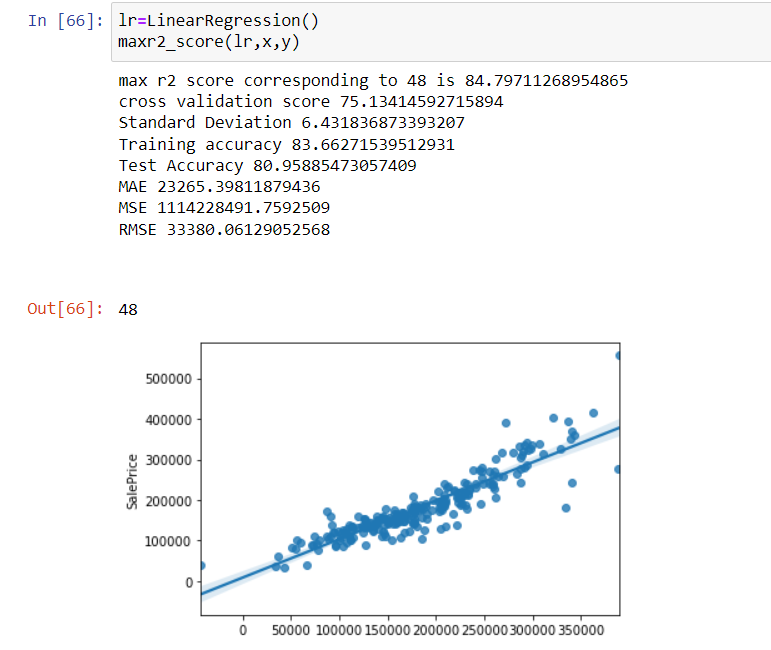
### Performance Metric:

For evaluating the performance of our models, we will use mean absolute error (MAE) and mean squared error (MSE). If the predicted value of the element, and the corresponding true value, then for all the elements, RMSE is calculated as:

In Linear Regressor model, The Root mean squared error value (RMSE) is high so we shouldcompare with more model.

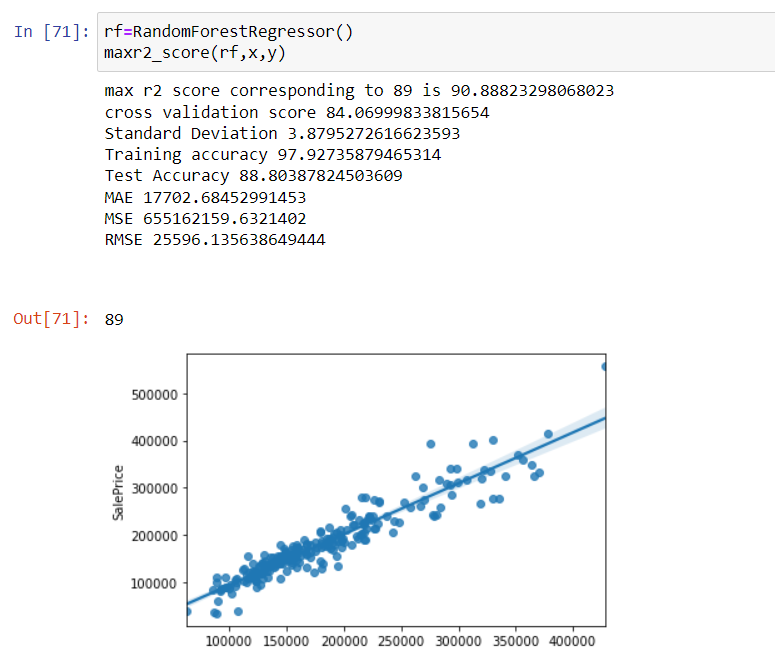


The Predict test and train values are calculated for Linear Regression model:



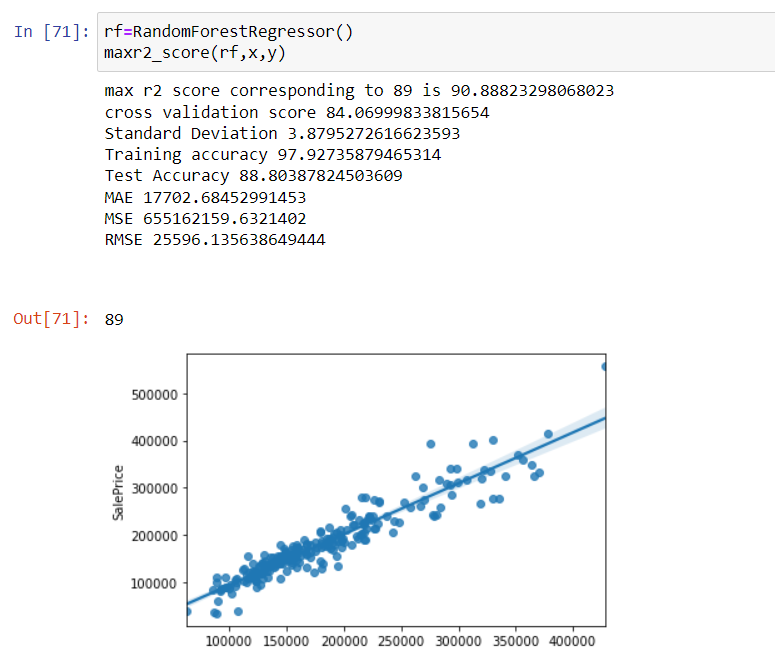
**HOUSE PRICE PREDICTION PROJECT pg. 23**

* + Using Scikit-Learn, we can build a model for Random Forest Regression Model



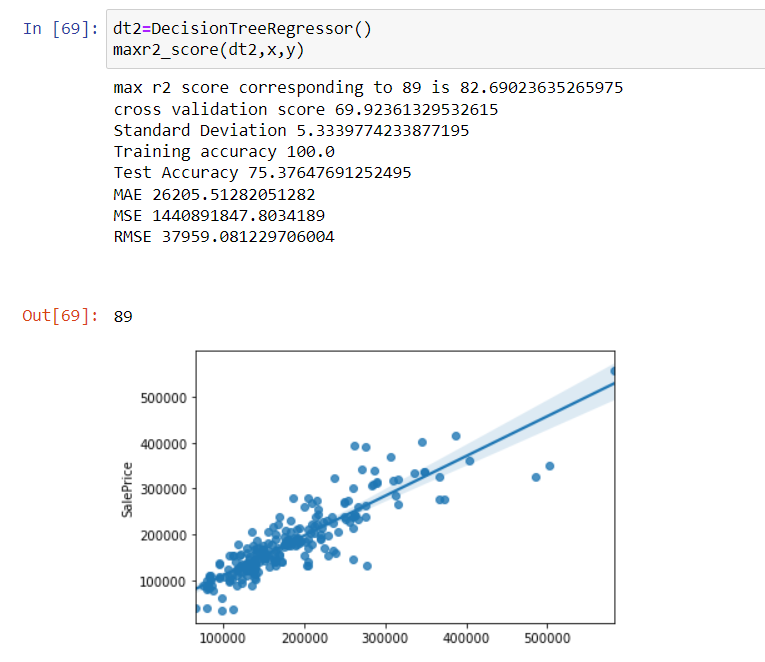
**HOUSE PRICE PREDICTION PROJECT pg. 24**

* + Using Scikit-Learn, we can build a model for Random Forest Regression Model

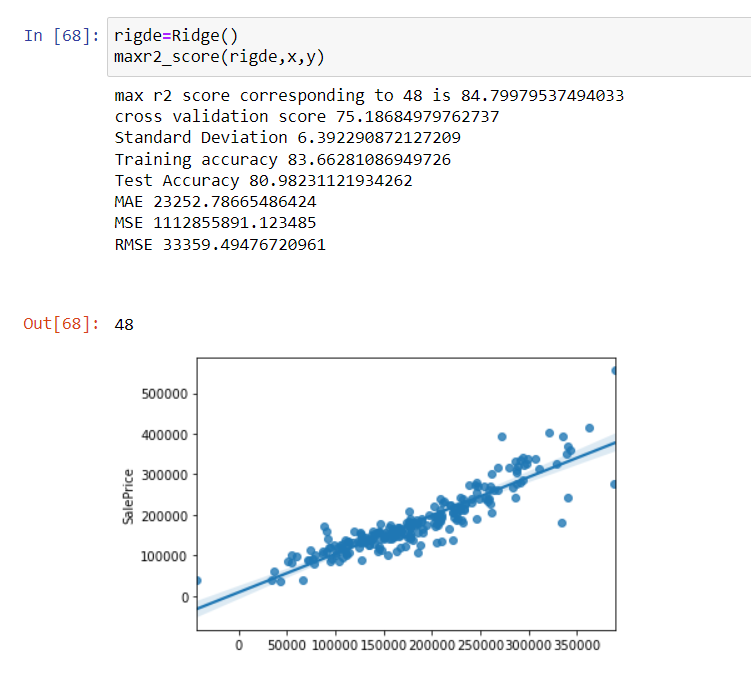


**HOUSE PRICE PREDICTION PROJECT pg. 25**

* + Using Scikit-Learn, we can build a model for Decision Tree Regression Model

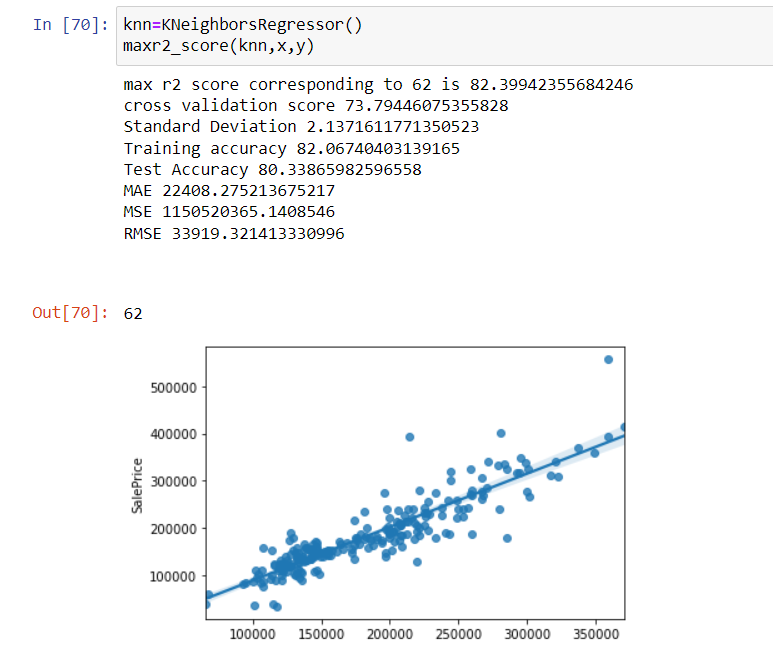


**HOUSE PRICE PREDICTION PROJECT pg. 26**

* + Using Scikit-Learn, we can build a model for Ridge Model. 

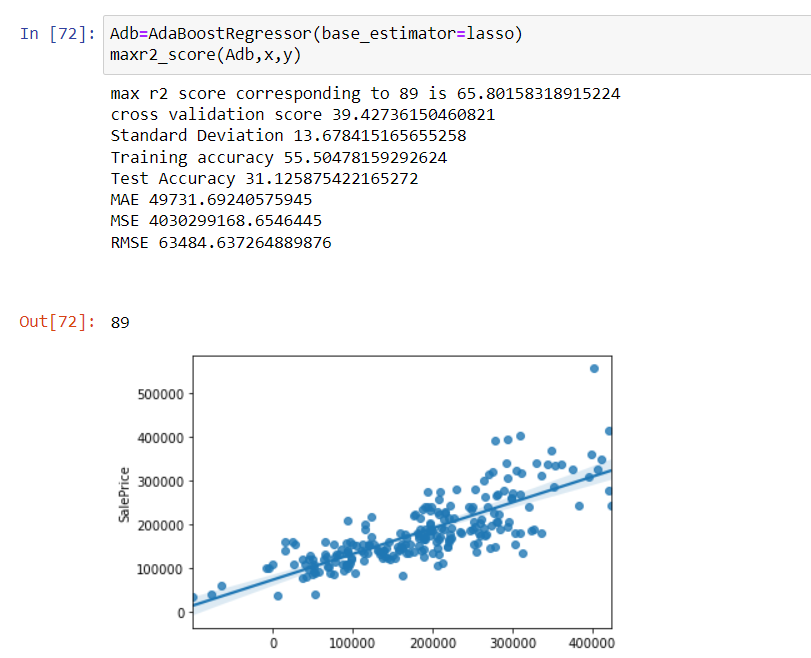
**HOUSE PRICE PREDICTION PROJECT pg. 27**

* Using Scikit-Learn, we can build a model for KNeighbor Regressor Model.



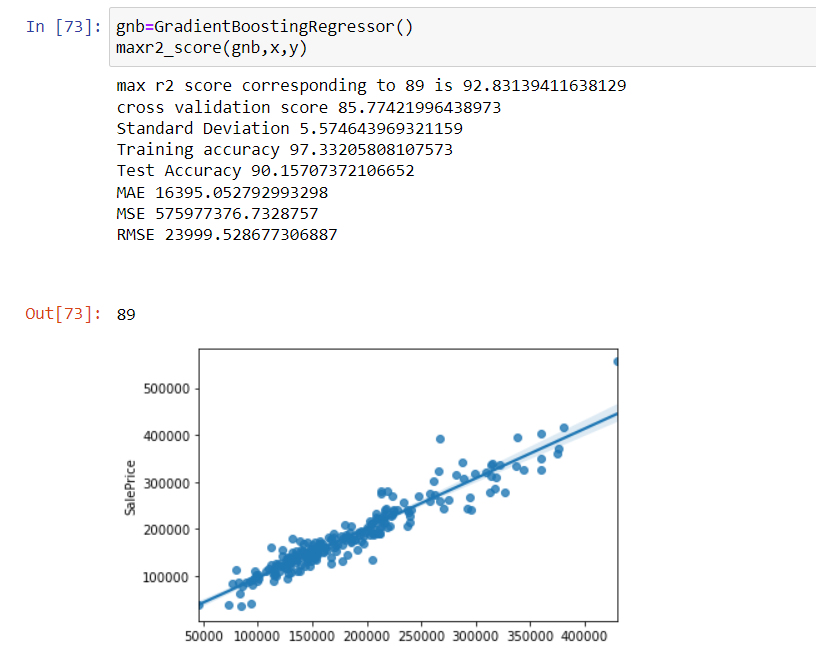
**HOUSE PRICE PREDICTION PROJECT pg. 28**

* Using Scikit-Learn, we can build a model for AdaBoost Regressor Model.



**HOUSE PRICE PREDICTION PROJECT pg. 29**

* Using Scikit-Learn, we can build a model for GradientBoost Regressor Model.



**HOUSE PRICE PREDICTION PROJECT pg. 30**

# Hyper Parameter Tuning:

Hyperparameters are crucial as they control the overall behaviour of a machine learning model. The ultimate goal is to find an optimal combination of hyperparameters that minimizes a predefined loss function to give better results.

The predict test value for both Random Forest regressor.

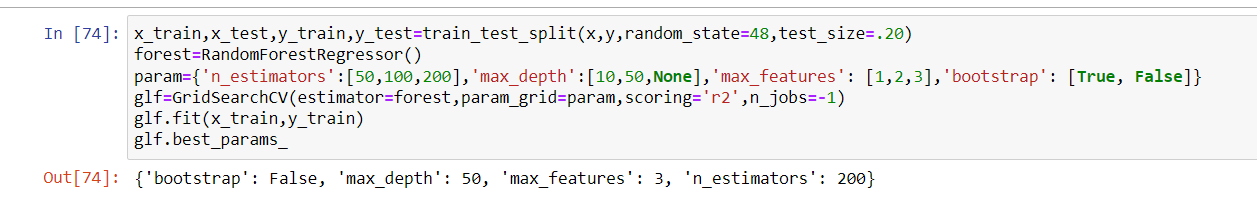
According to the performance of modelling approach, the RMSE value is low for Random Forest Regressor model but R2 score is similar for both Random Forest and XGBoost. Apply ensemble and regularization technique to determine the optimal values of Hyper Parameters.

#### Ensemble techniques:

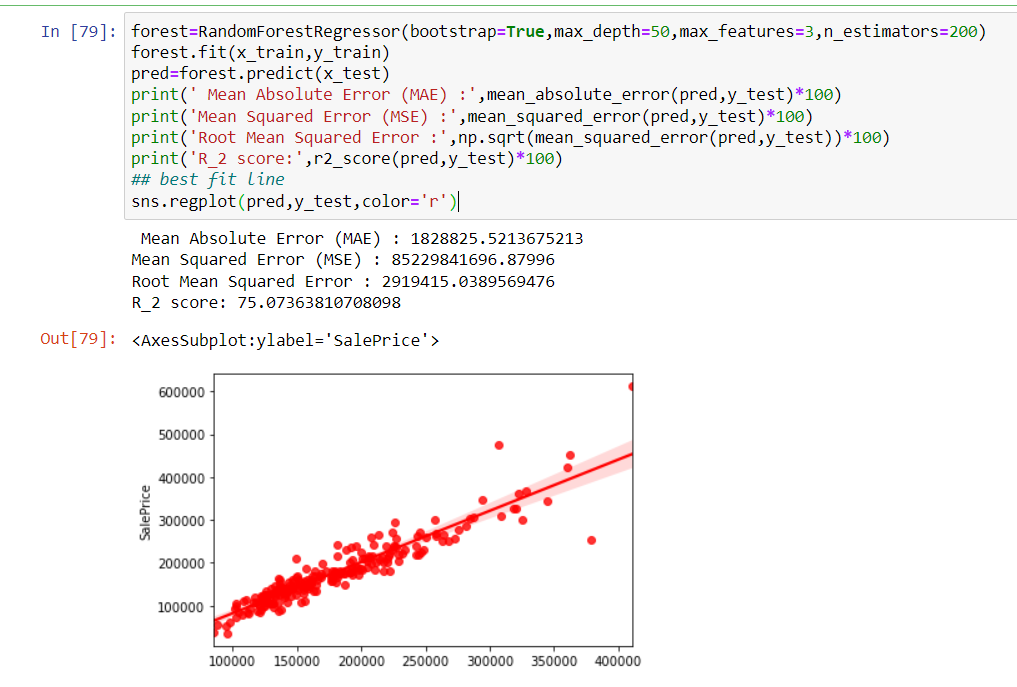
**Ensemble methods** is a machine learning **technique** that combines several base models in order to produce one optimal predictive model. To better understand this definition let’s takea step back into ultimate goal of machine learning and model building.

#### Hyperparameter Tuning for Random Forest Regressor:

Firstly, we will use GridSearchCV() to search for the best model parameters in a parameterspace provided by us. criterion, max features and random state.

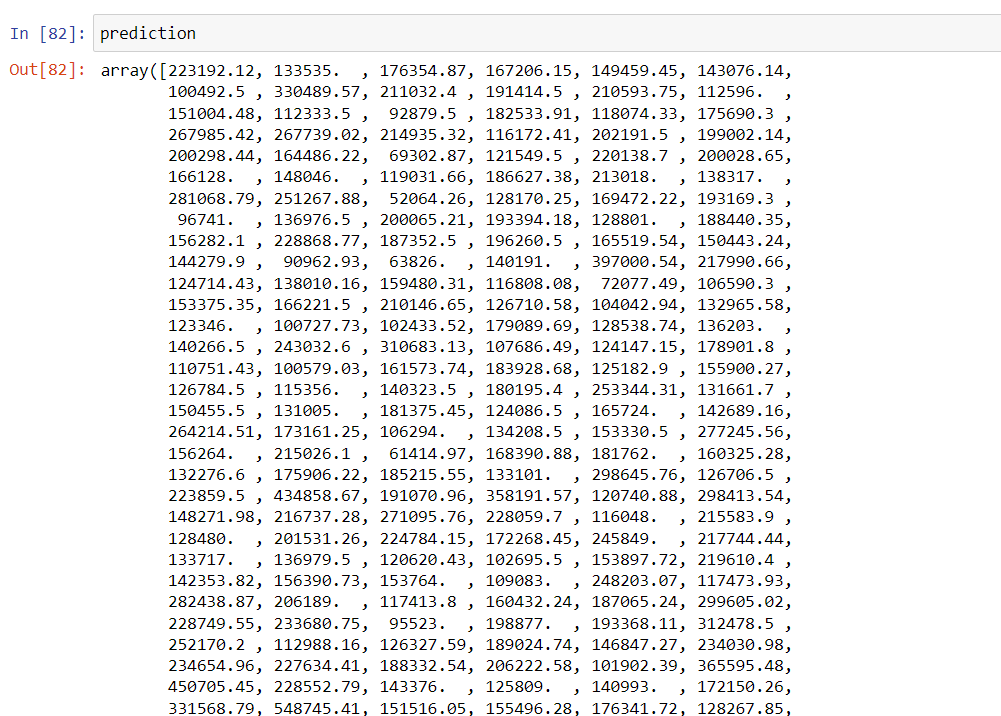


We defined the parameter space above using reasonable values for chosen parameters.



We defined the R2 score and Cross validation score of ensemble technique using chosen parameters. We are getting model accuracy and cross validation has 83.7% & 81.1% respectively.

**HOUSE PRICE PREDICTION PROJECT**

**HOUSE PRICE PREDICTION PROJECT pg. 31**

## Conclusion

In this paper, we built several regression models to predict the price of some house given some of the house features. We evaluated and compared each model to determine the one with highest performance. We also looked at how some models rank the features according to their importance. In this paper, we followed the data science process starting with getting the data, then cleaning and pre-processing the data, followed by exploring the data and building models, then evaluating the results and communicating them with visualizations.

As a recommendation, we advise to use this model (or a version of it trained with more recent data) by people who want to buy a house in the area covered by the dataset to have an idea about the actual price. The model can be used also with datasets that covered areas provided that they contain the same features. We also suggest that people take into consideration the features that were deemed as most important as seen in the previous section; this might help them estimate the house price better.

### Learning Outcomes of the Study in respect of Data Science:

* Obtain, clean/process, and transform data.
* Analyze and interpret data using an ethically responsible approach.
* Use appropriate models of analysis, assess the quality of input, derive insight from results, and investigate potential issues.
* Apply computing theory, languages, and algorithms, as well as mathematical and statistical models, and the principles of optimization to appropriately formulate anduse data analyses
* Formulate and use appropriate models of data analysis to solve hidden solutions to business-related challenges

HOUSE PRICE PREDICTION PROJECT pg. 42

### Limitations of this work and Scope for Future Work:

There are many things that can be tried to improve the models’ predictions. We can create and add more variables, try different models with different subset of features and/or rows, etc. Someof the ideas are listed below:

* + Combine the applicants with 1,2,3 or more dependents and make a new feature as discussed in the EDA part.
  + Make independent vs independent variable visualizations to discover some more patterns.
  + Arrive at the EMI using a better formula which may include interest rates aswell.
  + Try neural network using TensorFlow or PyTorch.

HOUSE PRICE PREDICTION PROJECT **pg. 32**

**pg. 33**

**HOUSE PRICE PREDICTION PROJECT pg. 37**

**HOUSE PRICE PREDICTION PROJECT pg. 38**

**HOUSE PRICE PREDICTION PROJECT pg. 39**

**HOUSE PRICE PREDICTION PROJECT pg. 40**

**HOUSE PRICE PREDICTION PROJECT pg. 41**

pg. 43