FLIP ROBO

Project - House Price Prediction

Submitted by:

Chaganti Sai Mahathi

ACKNOWLEDGMENT

References:

[1] P. Dent, "Animal Spirits – How Human Psychology Drives the Economy, and Why it Matters for Global Capitalism20106George A. Akerlof and Robert J. Shiller. Animal Spirits – How Human Psychology Drives the Economy, and Why it Matters for Global Capitalism. Woodstock: Princeton University Press 2010.", Journal of Property Investment & Finance, vol. 28, no. 4, pp. 312-313, 2010.

[2] A. Morone and E. Samanidou, "A simple note on herd behaviour", Journal of Evolutionary Economics, vol. 18, no. 5, pp. 639-646, 2007.

[3] Zillow, "Zillow Prize", Zillow Promotions, 2018. [Online]. Available: https://www.zillow.com/promo/zillow-prize/. [Accessed: 27- Jul- 2018].

[4] H. Hong, "Behavioural Finance: Introduction", European Financial Management, vol. 13, no. 3, pp. 389-393, 2007.

[5] L. Ackert, B. Church and N. Jayaraman, "Is There a Link Between Money Illusion and Homeowners' Expectations of Housing Prices?", Real Estate Economics, vol. 39, no. 2, pp. 251-275, 2011.

[6] K. Case and R. Shiller, "Is There a Bubble in the Housing Market?", Brookings Papers on Economic Activity, vol. 2003, no. 2, pp. 299-362, 2003.

[7] A. Ising, "Pompian, M. (2006): Behavioral Finance and Wealth Management – How to Build Optimal Portfolios That Account for Investor Biases", Financial Markets and Portfolio Management, vol. 21, no. 4, pp. 491-492, 2007.

[8]"Assumptions of Multiple Linear Regression - Statistics Solutions", Statistics Solutions, 2018. [Online]. Available:

http://www.statisticssolutions.com/assumptions-of-multiple-linear-regression.

[9] https://towardsdatascience.com

[10] "KNN Regression", Saedsayad.com, 2018. [Online]. Available:

http://www.saedsayad.com/k_nearest_neighbors_reg.htm. [Accessed: 21-April-2022].

[11]"Partial Least Squares (PLS)", Statsoft.com, 2018. [Online].

Available: http://www.statsoft.com/Textbook/Partial-Least-Squares.

[Accessed:21-April-2022].

[12] Regression Techniques:

https://www.researchgate.net/publication/340939997_House_Price_Prediction_Using_Various_Regression_Techniques[Accessed:20-April-2022].

[13]Prediction analysis: https://medium.com/codex/house-price-prediction-with-machine-learning-in-python

[Accessed:22-April-2022].

[14] S. Rosen, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition", Journal of Political Economy, vol. 82, no. 1, pp. 34-55, 1974.

[15] S. Song, "Modelling Worker Residence Distribution in the Los Angeles Region", *Urban Studies*, vol. 31, no. 9, pp. 1533-1544, 1994.

[16] G. Sirmans, D.Macpherson, & E. Zietz. The Composition of Hedonic Pricing Models. Journal of Real Estate Literature. vol. 13 no 1, pp. 1-30, 2005.

[17] A. Adair, S. McGreal, A. Smyth, J. Cooper and T. Ryley, "House Prices and Accessibility: The Testing of Relationships within the Belfast Urban Area", Housing Studies, vol. 15, no. 5, pp. 699-716, 2000.

[18] A. Can, "The Measurement of Neighborhood Dynamics in Urban House Prices", Economic Geography, vol. 66, no. 3, p. 254, 1990.

[19] I. Lake, A. Lovett, I. Bateman and B. Day, "Using GIS and large-scale digital data to implement hedonic pricing studies", *International Journal of Geographical Information Science*, vol. 14, no. 6, pp. 521-541, 2000.

[20] S. Orford, "Towards a Data-Rich Infrastructure for Housing-Market Research: Deriving Floor-Area Estimates for Individual Properties from Secondary Data Sources", Environment and Planning B: Planning and Design, vol. 37, no. 2, pp. 248-264, 2010.

[21] A. Nur, R. Ema, H. Taufiq and W. Firdaus, "Modeling House Price pp. 49-64, 1996

[22] Analytics: https://www.mindsmapped.com/regression-analysis-with-python/

[Accessed:19-April-2022]

[23] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." Journal of machine learning research 12.Oct (2011): 2825-2830.

INTRODUCTION

Buying a house is commonly the most important financial transaction for the average person. The fact that most prices are negotiated individually (unlike a stock exchange system) creates an environment that results in an inefficient system. A housing bubble cannot exist if individuals are making rational decisions.

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. The objective of this paper is to evaluate the performance of a stacked regression model compared to several sub models based on predicting house prices.

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.

Organization of this Study:

This paper contains an introduction, Domain objectives, methodology, Comparison construction, results & discussion and the conclusion.

The first phase of this study explores the introduction to this project and the Domain Objectives required to process this prediction.

The second part will focus on the methodology which details the Steps, Methods and materials used to analyse the data.

The Third section and the analytical stage will focus on constructing and tuning the sub algorithms to achieve an optimal performance. In this phase we will articulate the correlation between the sub algorithms predictions and select the optimal combination of sub algorithms to find the ideal stacked model.

The final side of this study sets forth the conclusions about the efficiency of the regression algorithm on house price analysis and makes recommendations for further use in production.

Domain Objectives:

- Regression models are used in this problem since the data to be predicted i.e., Sales price of houses is continuous.

- The machine learning models used to predict the price are 'Linear Regression' & 'Lasso regularisation', Ensemble approaches such as 'Random Forest Regressor', 'Decision tree regressor' and 'K-Neighbhor's Regressor'….and so on.

Linear Regression: Linear regression is the most common form of regression. Multi linear regression is used to explain the relationship between one continuous dependant variable and two or more explanatory variables. Linear regression has several assumptions such as; Regression residuals must be normally distributed, residuals are homoscedastic and approximately rectangular-shaped, Absence of multicollinearity is assumed in the model.

K Nearest Neighbour Regression: K-Neighbhor's Regressor is a non-parametric lazy learning algorithm. KNN uses a similarity measure such as the Euclidian distance to find a heuristically optimal number of K nearest neighbours based on Root mean square error to label new observations. The output is the property value for the object. This value is the average of the values of its k nearest neighbours. KNN is sensitive to high dimensional data and outliers.

Random Forest Regression: Random Forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression.

Decision Tree Regression: Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.

The branches/edges represent the result of the node and the nodes have either:

1. Conditions [Decision Nodes]
2. Result [End Nodes]

Here we use different techniques such as Data Analysis, Data mining and Data Visualisation, Data Pre-processing and model building is done for the prediction.

Literature Review:

A. **Economic point** of the housing market declares a plan that the housing market is not like any other market for goods or services. The housing market has the following traits:

- Every house serves as a commodity and also as an investment.
- A house generally builds up the highest portion of an individual’s net worth.
- Each individual house has a high cost of supply (Building and tax

costs).

- House is a highly durable and can last for several decades.
- These Houses are heterogeneous.
- They are immovable.
- Each House can be used as collateral against loans.
- A house can be sold for more money even for a second hand.

  This creates investment opportunities.

- House can generate rental income.

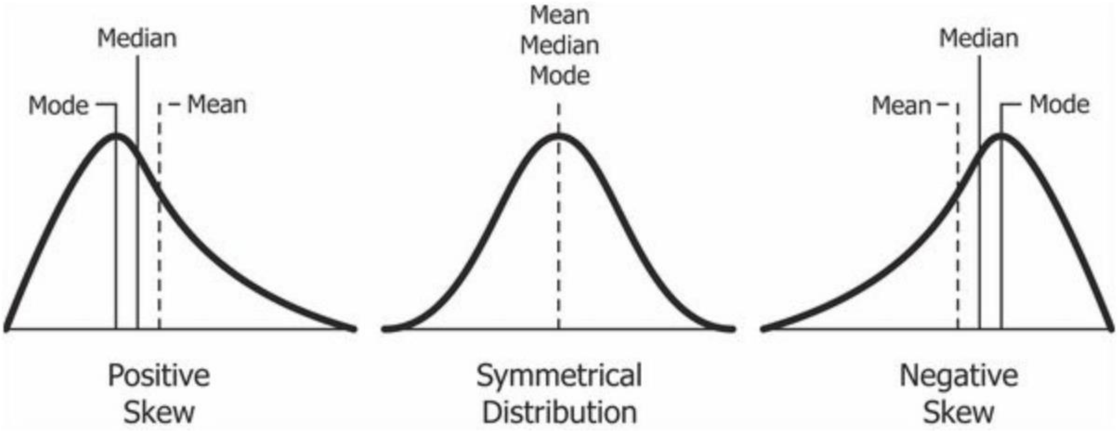
This says that the overall housing market is a group of many connected submarkets. It suggests that an increase in housing demand which increase the cost of housing is driven by biased on a speculation. Opposing the general theory of supply and demand, in housing, when prices rise demand increases and when price fall demand decreases. This can be explained by the psychological factors of buyers. According to the global housing bubble was driven by emotion rather that sound investment decisions. Overconfidence in decision making coupled with the ‘herd ‘like behaviour exasperated the housing crash that started in 2008. state that a bubble cannot exist if buyers are making rational decisions. This would indicate that buyers are not making rational decisions and are simple conforming to the trend.

B. **Hedonic price theory and its limitations**:

In most of the cases, price modelling is used to clarify the variation in price. This characteristic of a property is used as explanatory variables. The main concept in this field is the hedonic price theory. This theory can be traced back to the work complete by in 1966 and again in 1974. It is the main theory of reference to describe price in the housing market. In this concept, the price of a good, or in this instance a house, depends on its characteristics. Price theory uses the basic characteristics of a house such as number of bedroom and size, plenty research has been published on its limitations. Notable research focuses on the effect of location on price. Location factors include distance to work, transport systems, accessibility to schools and other amenities. One difficulty and drawback in using hedonic pricing models is that the results are location-specific and are difficult to generalize across different geographic regions. This suggests that that location is a key factor in determining price. But it is later discovered that buyers were willing to pay more for properties with quick accessibility to public transport. Despite of this argument housing is a multidimensional commodity separated into a package of attributes that vary in both quantity and quality. Accordingly, the hedonic housing price regression becomes an operational tool that functionally links housing expenditures to some measures of attributes of houses. In the study conducted by the articulated 125 hedonic models based on American house prices. They found that size related to the overall property and land had the best predictive utility.

C. **Skewness** is the measure of the asymmetry of an ideally symmetric probability distribution and is given by the third standardized moment. If that sounds way too complex, In simple words, skewness is the measure of how much the probability distribution of a random variable deviates from the normal distribution. Well, the normal distribution is the probability distribution without any skewness. You can look at the image below which shows symmetrical distribution that's basically a normal distribution and you can see that it is symmetrical on both sides of the dashed line. Apart from this, there are two types of skewness:

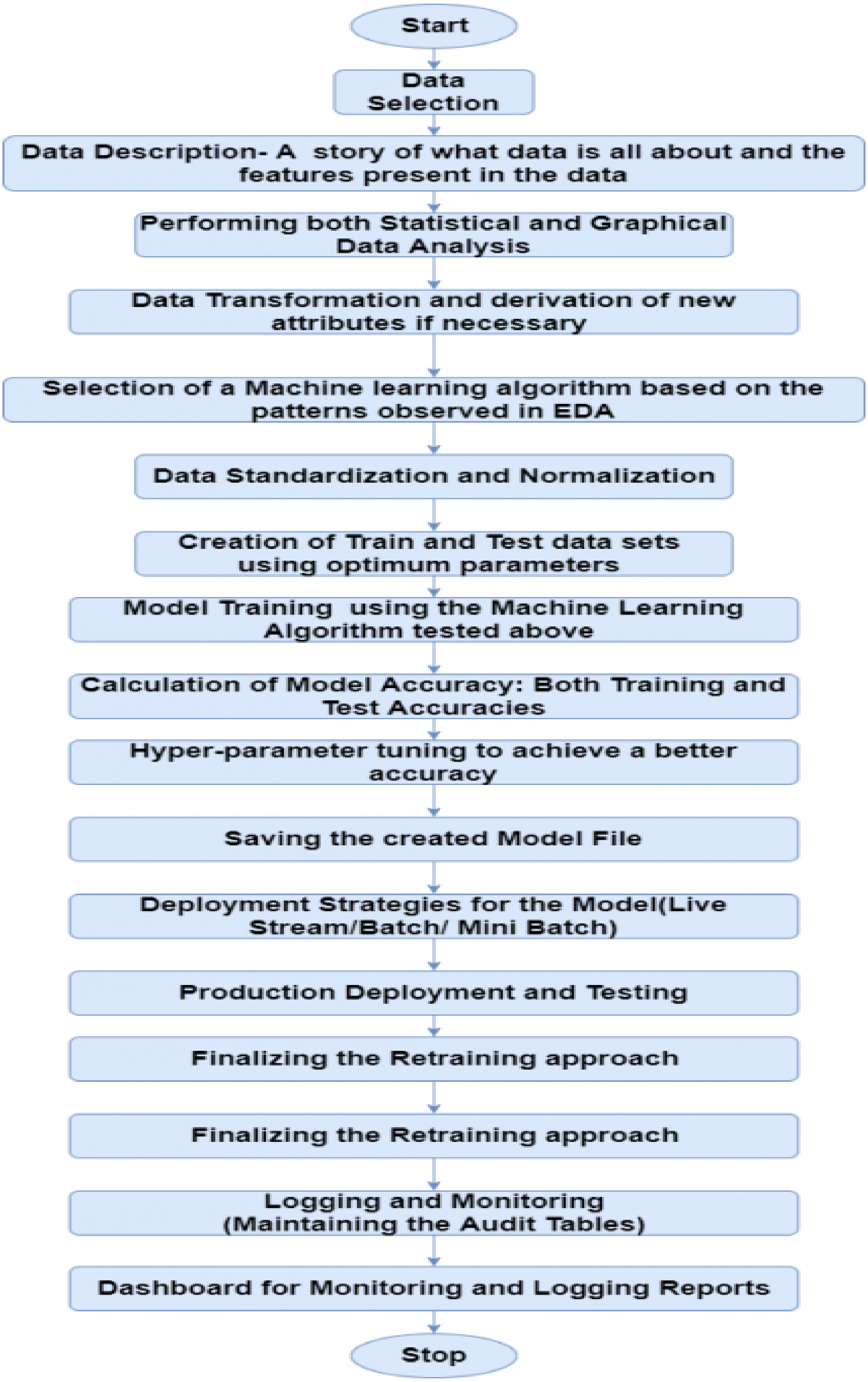
- Positive Skewness
- Negative Skewness



Firstly, linear models work on the assumption that the distribution of the independent variable and the target variable are similar. Therefore, knowing about the skewness of data helps us in creating better linear models.

D. **Regression modelling:**

Linear Regression is one of the most fundamental algorithms in the Machine Learning world. But before proceeding with the algorithm, let’s first discuss the life cycle of any machine learning model. This diagram explains the creation of a Machine Learning model from scratch and then taking the same model further with hyper-parameter tuning to increase its accuracy. A typical life cycle diagram for a machine learning model looks like:


Start
Data Selection
Data Description- A story of what data is all about and the features present in the data
Performing both Statistical and Graphical Data Analysis
Data Transformation and derivation of new attributes if necessary
Selection of a Machine learning algorithm based on the patterns observed in EDA
Data Standardization and Normalization
Creation of Train and Test data sets using optimum parameters
Model Training using the Machine Learning Algorithm tested above
Calculation of Model Accuracy: Both Training and Test Accuracies
Hyper-parameter tuning to achieve a better accuracy
Saving the created Model File
Deployment Strategies for the Model(Live Stream/Batch/ Mini Batch)
Production Deployment and Testing
Finalizing the Retraining approach
Finalizing the Retraining approach
Logging and Monitoring (Maintaining the Audit Tables)
Dashboard for Monitoring and Logging Reports
Stop

The benefits of using Regression analysis are as follows:

- It shows the significant relationships between the Label (dependent variable) and the features (independent variable).
- It shows the extent of impact of multiple independent variables on the dependent variable.
- It can also measure these effects even if the variables are on a different scale.

These features enable the data scientists to find the best set of independent variables for predictions.

A linear regression model consists of:

- Discreet/continuous independent variables
- A best-fit regression line
- Continuous dependent variable. The equation of the Linear Regression is **Y=a+b*X + e** where, a is the intercept, b is the slope of the line, and e is the error term. The equation above is used to predict the value of the target based on the given predictors.

Primary Purpose of this Project:

- 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.
- The main object of this project is to model the 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.

Methods and Materials:

In the following paragraphs we describe the materials (Data set and features) and methods (software) used to develop this comparison.

Software:

This analysis was complete using the Python programming language. The Python package was used to streamline the construction of the models. This allowed us to split the data

into test and train sets, pre-process, cross validate and tune the algorithms. We then used the Ensemble package for the better prediction models.

Data set and exploratory analysis:

Data contains 1460 entries each having 81 variables. Here there are 2 datasets provided one consist of 1168 rows and 81 columns which is training data set and other with 292 rows and 80 columns as testing data set .

The Target variable is Sale Price which is to be predicted. Here, prediction is made by the price of a house with respect to feature variable conditions.

Data and its Description of feature variables:

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

@MSZoning: Identifies the general zoning classification of the sale.

@LotFrontage: Linear feet of street connected to property

@LotArea: Lot size in square feet

@Street: Type of road access to property

@Alley: Type of alley access to property

@LotShape: General shape of property

@LandContour: Flatness of the property

@Utilities: Type of utilities available

@LotConfig: Lot configuration

@LandSlope: Slope of property

@Neighborhood: Physical locations within Ames city limits

@Condition1: Proximity to various conditions

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

@BldgType: Type of dwelling

@HouseStyle: Style of dwelling

@OverallQual: Rates the overall material and finish of the house

@OverallCond: Rates the overall condition of the house

@YearBuilt: Original construction date

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

@RoofStyle: Type of roof

@RoofMatl: Roof material

@Exterior1st: Exterior covering on house

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

@MasVnrType: Masonry veneer type

@MasVnrArea: Masonry veneer area in square feet

@ExterQual: Evaluates the quality of the material on the exterior

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

@Foundation: Type of foundation

@BsmtQual: Evaluates the height of the basement

@BsmtCond: Evaluates the general condition of the basement

@BsmtExposure: Refers to walkout or garden level walls

@BsmtFinType1: Rating of basement finished area

@BsmtFinSF1: Type 1 finished square feet

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

@BsmtFinSF2: Type 2 finished square feet

@BsmtUnfSF: Unfinished square feet of basement area

@TotalBsmtSF: Total square feet of basement area

@Heating: Type of heating

@HeatingQC: Heating quality and condition

@CentralAir: Central air conditioning

@Electrical: Electrical system

@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

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

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

@Fireplaces: Number of fireplaces

@FireplaceQu: Fireplace quality

@GarageType: Garage location

@GarageYrBlt: Year garage was built

@GarageFinish: Interior finish of the garage

@GarageCars: Size of garage in car capacity

@GarageArea: Size of garage in square feet

@GarageQual: Garage quality

@GarageCond: Garage condition

@PavedDrive: Paved driveway

@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

@Fence: Fence quality

@MiscFeature: Miscellaneous feature not covered in other categories

@MiscVal: $Value of miscellaneous feature

@MoSold: Month Sold (MM)

@YrSold: Year Sold (YYYY)

@SaleType: Type of sale

@SaleCondition: Condition of sale

All the above features are responsible for the house sale market prediction.

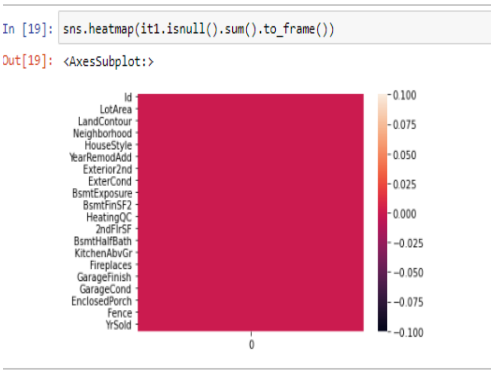
A} Removing null values:

Imputing numerical valued variables to remove null values present in both the test and train datasets using Iterative imputer.

Filling all the other non-numeric variables null values using lambda function and ‘fillna’ method.

Checking the null count using a heat map of training and testing data respectively:

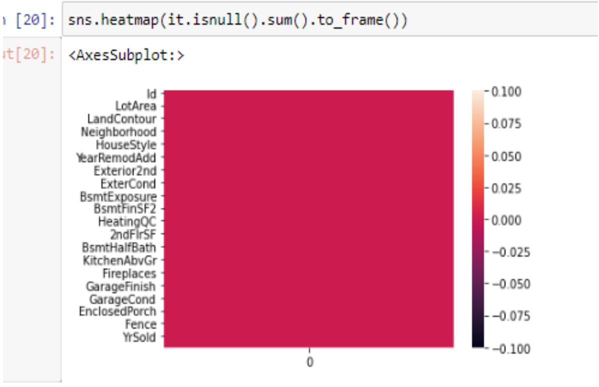
Testing Data:



```
In [19]: sns.heatmap(it1.isnull().sum().to_frame())

Out[19]: <AxesSubplot:>
```

Training Data:



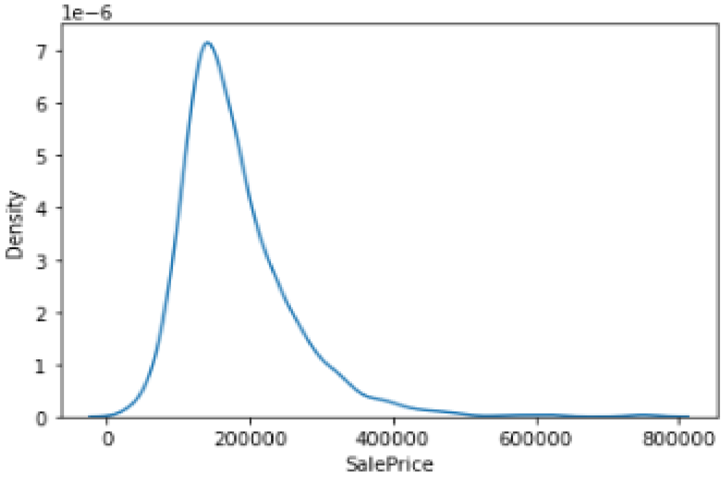
```
[20]: sns.heatmap(it.isnull().sum().to_frame())

t[20]: <AxesSubplot:>
```

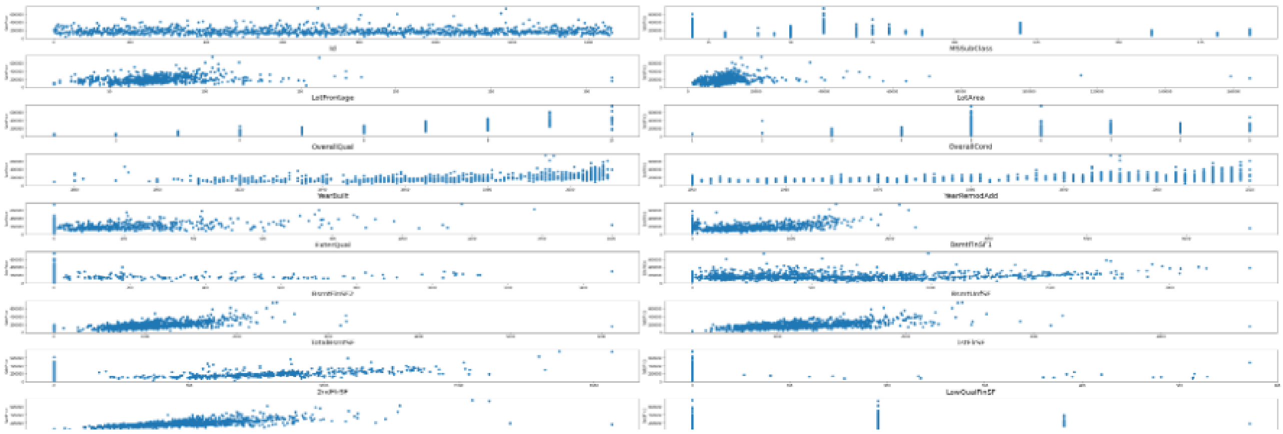
B} Data Visualisation and analysis:

The target variable Sale Price is shown in the dist. plot of seaborn to get the better visualisation of the housing price.

```
<AxesSubplot:xlabel='SalePrice', ylabel='Density'>
```



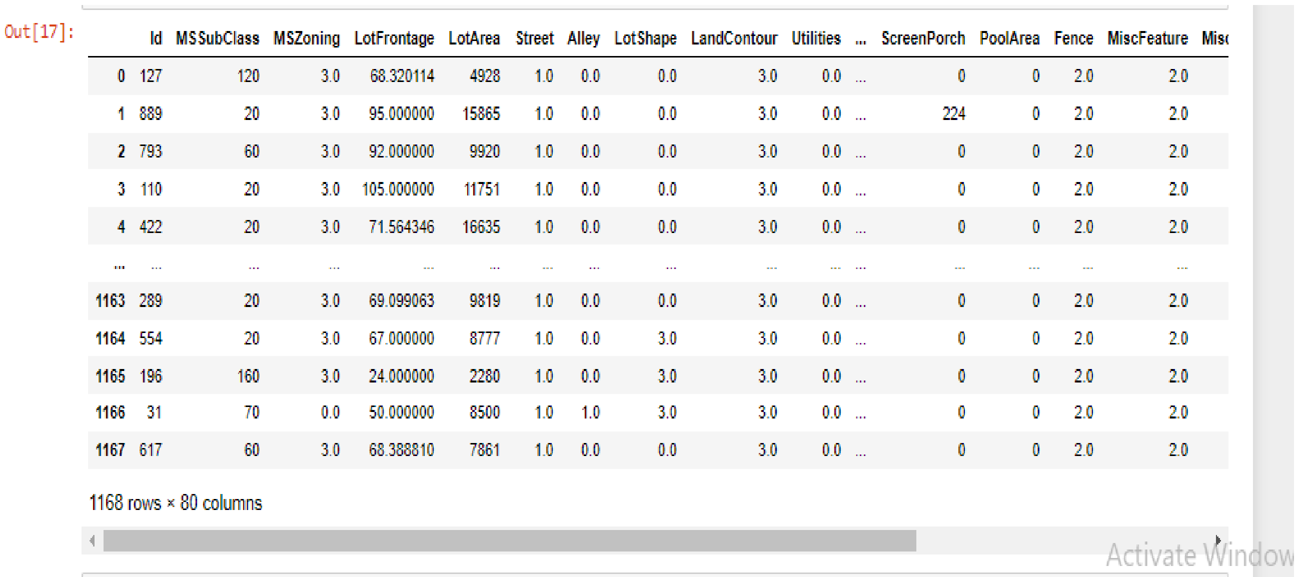
For better understanding of the feature variables, we then plotted a scatter plot with respect to the sale price variable of trained data.



Since there is huge data it is difficult to analyse as is but seeing all the plots at once can say that the features with respect to target variable are almost linearly spread across the plot.

C} Encoding the data:

TO convert all object type data to numerical of both training and testing data we use Ordinal encoder.



Out[17]:

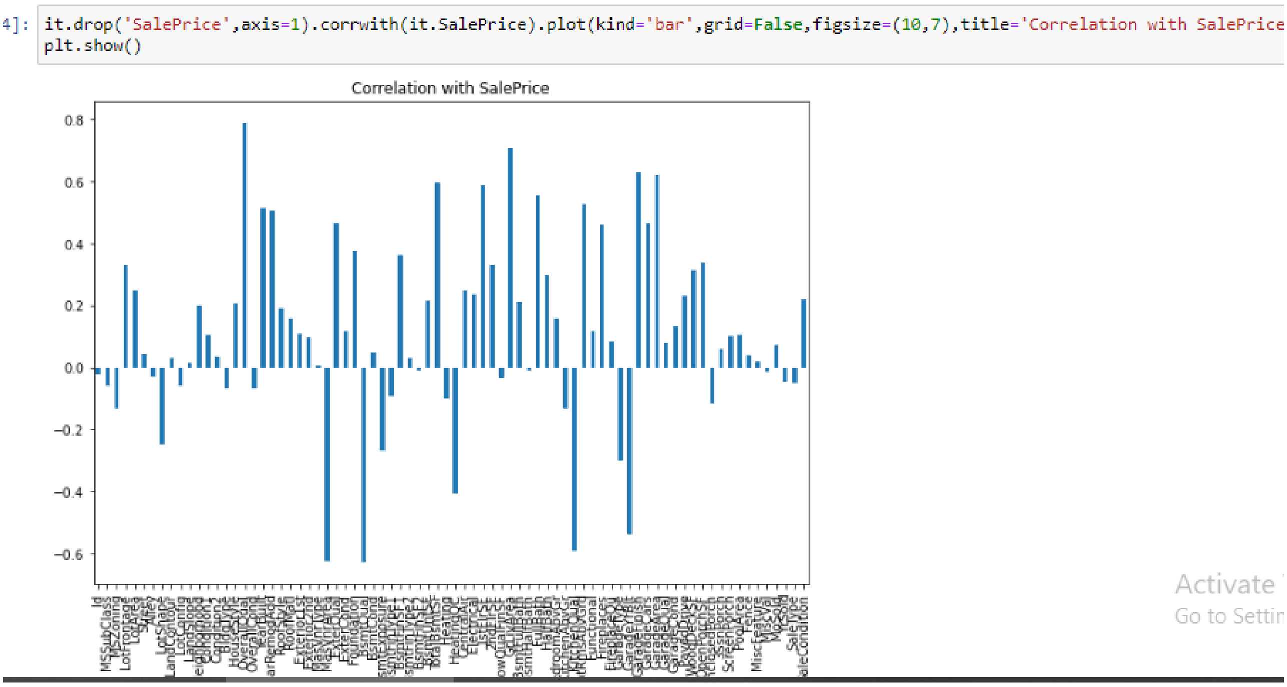
	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	ScreenPorch	PoolArea	Fence	MiscFeature	Misc
0	127	120	3.0	68.320114	4928	1.0	0.0	0.0	3.0	0.0	...	0	0	2.0	2.0	
1	889	20	3.0	95.000000	15865	1.0	0.0	0.0	3.0	0.0	...	224	0	2.0	2.0	
2	793	60	3.0	92.000000	9920	1.0	0.0	0.0	3.0	0.0	...	0	0	2.0	2.0	
3	110	20	3.0	105.000000	11751	1.0	0.0	0.0	3.0	0.0	...	0	0	2.0	2.0	
4	422	20	3.0	71.564346	16635	1.0	0.0	0.0	3.0	0.0	...	0	0	2.0	2.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1163	289	20	3.0	69.099063	9819	1.0	0.0	0.0	3.0	0.0	...	0	0	2.0	2.0	
1164	554	20	3.0	67.000000	8777	1.0	0.0	3.0	3.0	0.0	...	0	0	2.0	2.0	
1165	196	160	3.0	24.000000	2280	1.0	0.0	3.0	3.0	0.0	...	0	0	2.0	2.0	
1166	31	70	0.0	50.000000	8500	1.0	1.0	3.0	3.0	0.0	...	0	0	2.0	2.0	
1167	617	60	3.0	68.388810	7861	1.0	0.0	0.0	3.0	0.0	...	0	0	2.0	2.0	

1168 rows × 80 columns

D}Data Pre-Processing:

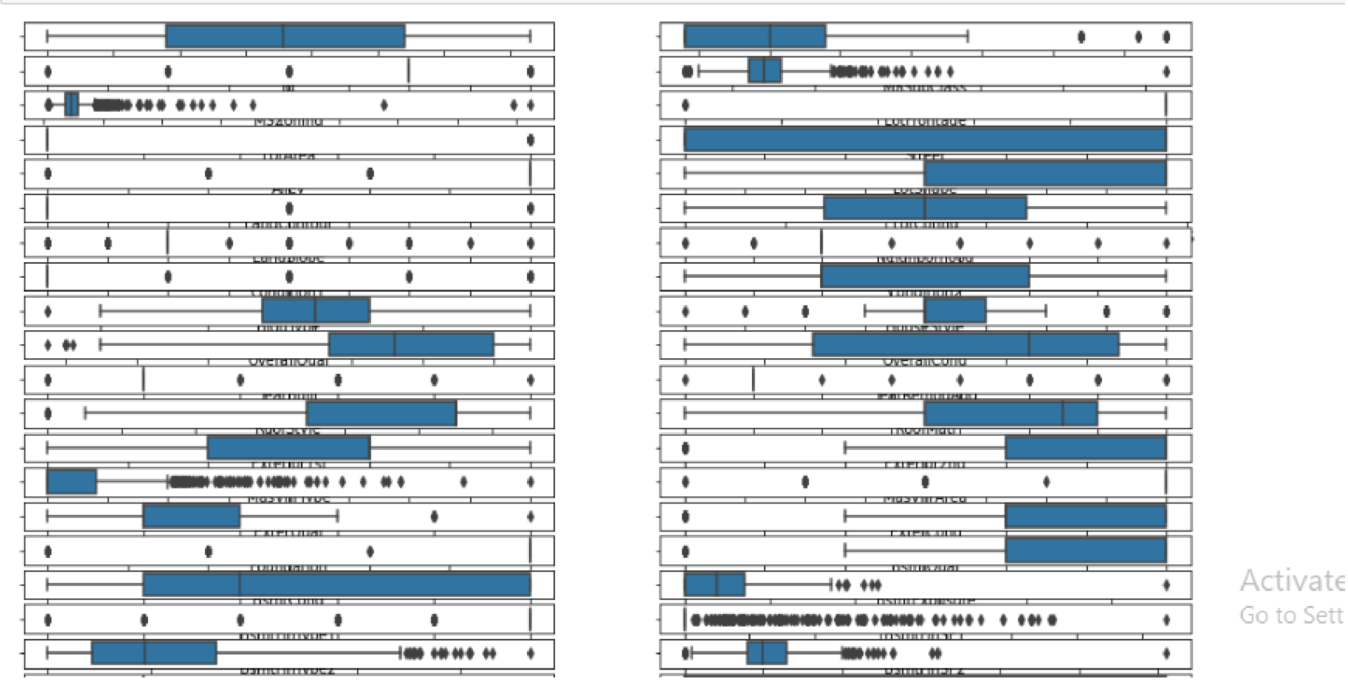
The first step is to Drop all the unwanted column present in the dataset.

Now, since the data sets are ready to process, check the correlation of the training dataset:



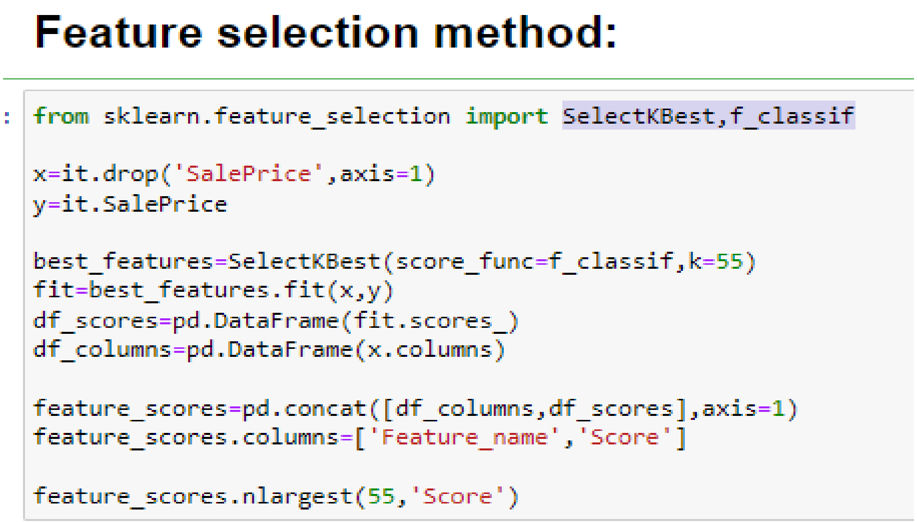
```
4]: it.drop('SalePrice',axis=1).corrwith(it.SalePrice).plot(kind='bar',grid=False,figsize=(10,7),title='Correlation with SalePrice
    plt.show()
```

Now lets also check the skewness and outliers if exists lets try to remove using some commonly used techniques.



Removing outliers and skewness is a very difficult process in this case so it is better to go with Feature selection process.

In Feature selection method, we selected top 55 features out of 79 features present in the dataset by importing SelectKBest, f_classif methods.



Feature selection method:

```
from sklearn.feature_selection import SelectKBest,f_classif

x=it.drop('SalePrice',axis=1)
y=it.SalePrice

best_features=SelectKBest(score_func=f_classif,k=55)
fit=best_features.fit(x,y)
df_scores=pd.DataFrame(fit.scores_)
df_columns=pd.DataFrame(x.columns)

feature_scores=pd.concat([df_columns,df_scores],axis=1)
feature_scores.columns=['Feature_name','Score']

feature_scores.nlargest(55,'Score')
```

Now once the best features are declared we cannot say that we avoided any outliers or skewness completely but now by considering these features let’s create a new training and testing data frames.

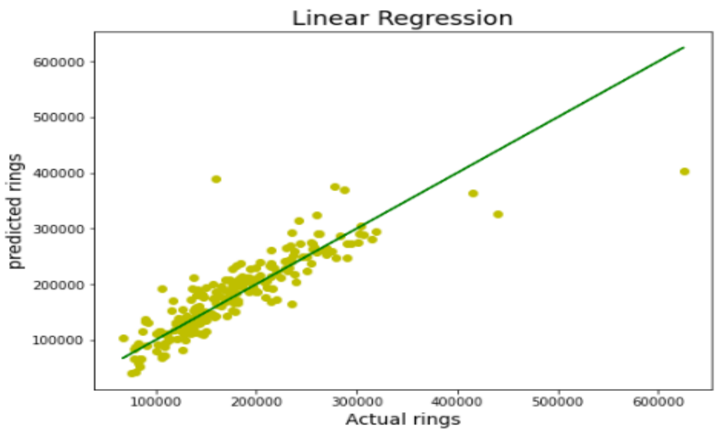
Model Building:

Now, once the data is pre processed and ready for the building, we first Train the data using various models and then use this trained data for testing or predicting the testing model.

First split the feature variables as ‘x’ data frame and ‘y’ as target. Then, Transform the ‘x’ variable using power transformation to remove any skewness present in the data and then normalise the data using Standard Scaler.

Next create a train test split of random state given as 41and test size as 0.20 and then apply linear regression model to test the data the r2 score of the predicted d ata is given as **0.7423741774561421.**

Linear Regression Curve:

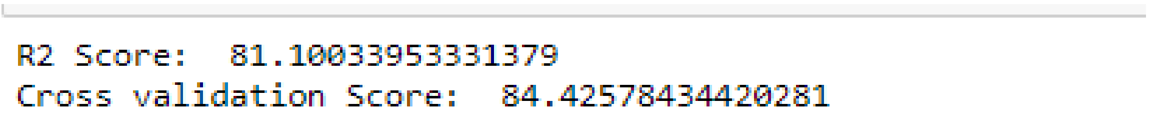


Once the model is linearly regressed then use lasso regularisation technique to get the maximum best output-0.7425260295759903
Using the best cross validation score we get the output for this model as

```
0.7875568967830889
```

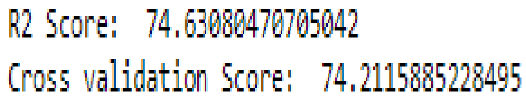
Now we use other Ensemble approaches and further using hyper parameter tuning finding the best parameters gets the better results of

Random Forest Regression:



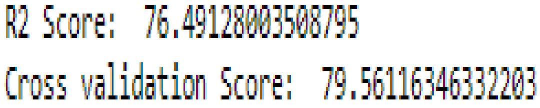
```
R2 Score:  81.10033953331379
Cross validation Score:  84.42578434420281
```

Decision Tree Regression:



```
R2 Score:  74.63080470705042
Cross validation Score:  74.2115885228495
```

K-Neighbors Regression:



```
R2 Score:  76.49128003508795
Cross validation Score:  79.5611634633203
```

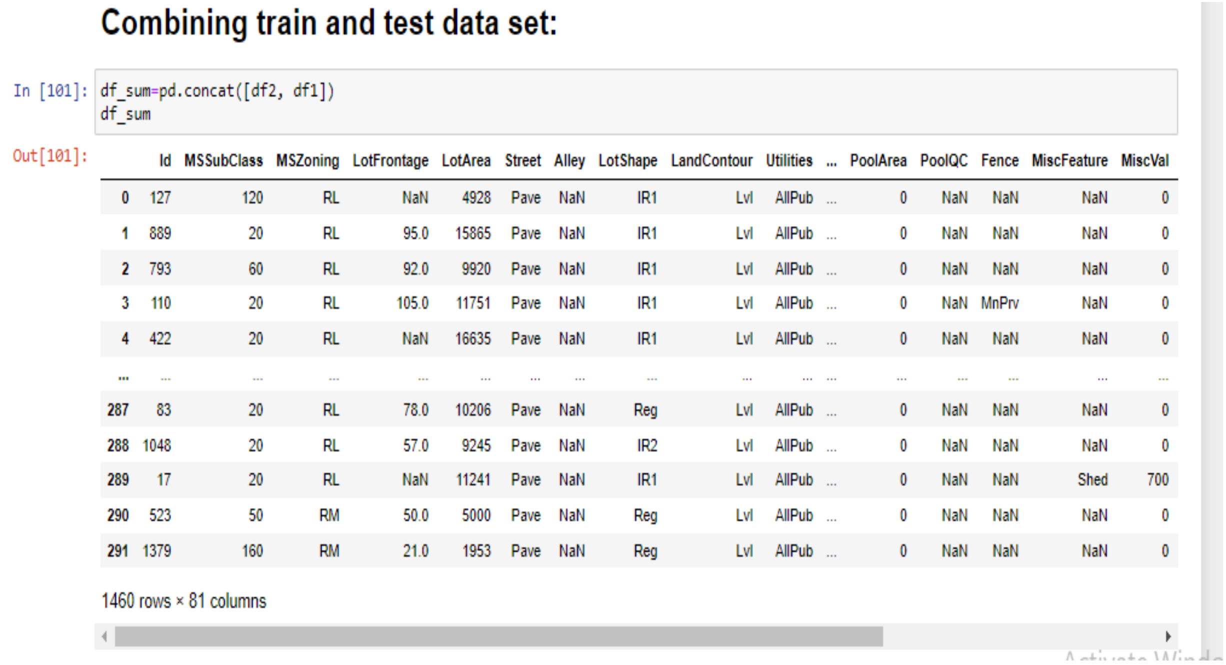
Out of all models Random Forest regression is having R2 Score: 81.10033953331379 and Cross validation Score: 84.42578434420281. So, this model is finalised and further saved.

Saving the model and Loading the data is the further most step, now is that my model is trained it goes to testing phase

Predicting the SalePrice for tested dataset using trained model:

Prediction of the modified testing data using the trained model is done using the trained random forest model and join with the test data.

Now once the model predicted join the test dataset in which sale price predicted column added to the train dataset. Hence the complete data with 1460 rows and 81 columns is defined.



Combining train and test data set:

In [101]:

```
df_sum=pd.concat([df2, df1])
df_sum
```

Out[101]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal
0	127	120	RL	NaN	4928	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
1	889	20	RL	95.0	15865	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
2	793	60	RL	92.0	9920	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
3	110	20	RL	105.0	11751	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	MnPrv	NaN	0
4	422	20	RL	NaN	16635	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	NaN	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
287	83	20	RL	78.0	10206	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0
288	1048	20	RL	57.0	9245	Pave	NaN	IR2	Lvl	AllPub	...	0	NaN	NaN	NaN	0
289	17	20	RL	NaN	11241	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN	NaN	Shed	700
290	523	50	RM	50.0	5000	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0
291	1379	160	RM	21.0	1953	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN	NaN	NaN	0

1460 rows × 81 columns

CONCLUSION & FUTURE WORK:

The above trained regression model with 1460 rows and 81 columns is able to predict only 81% with cross validation around 84% which can be further processed with maximum results using several techniques to remove outliers completely.

This data set has many limitations. The largest limitation is that we have no information about potential buyers and the environment of the sale. Factors such as auctions can have an influence on the price of a house due to bidding wars and ego. The data was gathered over a period of one year, so it does not capture much seasonality and does not consider economic factors.

An interesting area for future work would be to examine buyer related data with house characteristics.