House Sale Price Prediction

Raju Ranjan
*rajur.ug*18.*ec@nitp.ac.in*
B.Tech., Department of
Electronics and Communication Engineering
National Institute of Technology Patna, India

Shruti Sinha
*shrutis.ug*18.*ec@nitp.ac.in*
B.Tech., Department of
Electronics and Communication Engineering
National Institute of Technology Patna, India

Abstract

Determining the sale price of the house is very important nowadays as the priceof the land and price of the house increases every year. So our future generationneeds a simple technique to predict the house price in future. The price of househelps the buyer to know the cost price of the house and also the right time tobuy it. The right price of the house helps the customer to elect the house andgo for the bidding of that house. There are several factors that affect the priceof the house such as the physical condition, location, landmark etc. This paperuses various regression techniques to predict the house price such as Ridge, Lasso,ElasticNet regression techniques

Keywords -

House Price; Data Exploration; Sale Price; Target Variable; Linear Regression; Feature Engineering ,etc.

1 Introduction

Purchasing a house is a big decision in a person's life and needs a considerable amount of thought and research. One would like to buy a house at the best rate and minimum risk and would like it to be the best investment for the future. Various online websites, real estate agents and realtors try to guide home buyers by letting them compare different houses available for purchase. In this study, several methods of prediction were compared to finding the best predicted results in determining a house's selling price compared to the actual price.

This paper brings the latest research on regression technique that can be used for house sale price prediction such as Linear regression,Gradient boosting and hybrid regression. As the initial house price prediction were challenging and require some best method to get accurate prediction. Data quality is a key factor to predict the house prices and missing features are a difficult aspect to handle in machine learning models let alone house prediction model. Therefore, feature engineering becomes an important method for creating models which will give better accuracy. In general, the value of the property increases over time and its valued value must be calculated. During the sale of property or while applying for the loan and the marketability of the property, this valued value is required. The professional evaluators determine these valued values. However, the disadvantage of this practice is that these evaluators could be biased because buyers, sellers or mortgages have bestowed interest. We therefore need an automated model of prediction that can help to predict property values without bias. This automated model can help first - time buyers and less experienced customers to see if property rates are overrated or underrated.

2 Data Analysis

The dataset contains two types of variables:

2.1 Numeric Variables

There are 10 numerical features that are relevant. , which "identifies the type of residence involve d in the sale,". There are 17 numerical features, of the following types:

ID, No of Bedrooms , Sale Price, No of Bathrooms, Flat Area (in Sqft), Lot Area (in Sqft), Area of the House from Basement (in Sqft), Basement Area (in Sqft), No of Floors, Overall Grade , Age of House (in Years), Latitude, Longitude, Renovated Year, Zipcode , Living Area after Renovation (in Sqft), Lot Area after Renovation (in Sqft).

Condition and quality: Most of the variables dealing with the apartment's actual physical space are positively skewed—which makes sense because people tend to live in smaller homes / apartments apart from the extremely rich.

2.2 Categorical Variables

A categorical variable is a variable that can take on one of a limited, and usually fixed, number of possible values, assigning each individual or other unit of observation to a particular group or nominal category on the basis of some qualitative property. There are 4 Categorical features that are relevant and used in building the model.

There are 21 Categorical features, of the following types: Condition of the House, Ever Renovated, Waterfront View .

3 Methodology

The design approach involves pre-processing of data, creative feature engineering and the regression model such as ridge regression, Gradient boosting, Linear regression and Hybrid regression.

3.1 Data Exploration and Preprocessing

Figure 1. Data Preprocessing

3.1.1 Target Variable

The questions one should usually find answers to while exploring the target variable

- What information does this variable represent?
- What is the data type for this variable?
- Do the few sample values that we can 'eyeball' (i.e. skim through) make sense?
- What are the minimum, maximum, mean, median values for this variable and whether these make sense and what we can infer from them?
- Does the target variable contain any Outliers that we need to treat?
- Are there any missing values in the target variable that we need to treat?
- What is the distribution of the values of this variable over its range? Are the values uniformly or normally distributed or is data skewed towards lower values (i.e. contains more of lower values) or towards higher values?

We have the Target Variable as 'Sale Price'.

1. We find that our target variable is Sale Price in dollars at which the house was sold and it is of numeric type data and is not normally distributed as its mean and median are not equal
2. **Eyeballing** - Skimming through the values of the variable of interest and see whether you notice any patterns or any anormal by just looking at the data. For that find the descriptive analysis of the data. We got that mean is greater than median, which indicates it may contain outliers of high values or could be skewed towards lower values.
3. **Outliers**: A data point that is distant from the rest of the observations. How to identify outliers mathematically: Creating a Box and Whisker plot in python to identify the Outliers.

$$\text{lower\_limit} = Q1 - (1.5*IQR)$$
$$\text{upper\_limit} = Q3 + (1.5*IQR)$$

where IQR is Inter Quartile Range.

Any data point is higher than upper_limit and lower than lower_limit to be considered as Outliers.

Treating Outliers :

(a) Deletion : Remove the Outliers , might cause data loss.

(b) Capping or Imputing : Replacing the Outliers with descriptive analysis.

(c) Data Transformation : Taking log or square, cube root of outliers

(d) Binning : Different bins are formed based on the values.

We will use Imputing here , so that we don't cause any essential data loss.we will have a limit_imputer function to get the allowed range for the data variable.
Resultant dataset will be named as this: Raw_Housing_Prices1.csv.

4. **Identify and Treating Missing values** : Whenever the values are not present or not available for a variable in a particular observation, the variable is said to contain missing values

The missing values can reduce the performance of the model.

Any row which has a missing value either for the target variable or any of the independent variables cannot be used for building the model

Plot histogram for the Target Variable and check the info to identify the missing values.

Treating missing values:

(a) Deletion : delete the entire row.

(b) Capping or Imputing : not recommended , predictive should not change target variable

– So we will use deletion in this case

3.1.2 Independent Variable

1. **Treating missing values:**

   (a) Deletion : delete the entire row.

   (b) Capping or Imputing : Using mean or median for the continuous variables , mode for the object variables.

   So we will use imputing in this case.
   We will use the sklearn library to do this which uses imputer function and fit Transform.
   Steps involved in "fit-Transform" function:

   - Fitting phase - The function calculates the median value with respect to every column that we passed in as a parameters and stores it.
   - Transformation phase - The function does the actual action of locating the missing values and imputing then using the median strategy.

   Now the modified i.e. Resultant dataset will be Raw_Housing_Prices3.csv

2. **Perform Variable Transformations :** Refers to the process of making changes to a variable in a way that it becomes more useful and meaningful for analysis and modeling and modeling purposes.

   Reasons are Outliers treatment and when a variable does not contain its best possible way.

   (a) Identifying the numerical variables.
   It makes sense to convert Zipcode variable data type into categorical variables where
   - Sale Price in each Zipcode can be analysed
   - Variations across Zipcode can be seen.

   (b) No. of time visited variable should be changed in numerical values.
   - Find all unique values
   - Replace them with their corresponding numerical values.

   Now the modified i.e. Resultant dataset will be Transformed_Housing_Prices.csv

3. **Correlation :** Measure of dependencies between two variables, how close two variables are to have a linear relationship with each other.

   - Usually select only that variable which has high correlation with the data variable.
   i.e.correlation >0.5
   - If two independent variables are highly correlated and also correlated with dependent variables then it may result in a poor model.

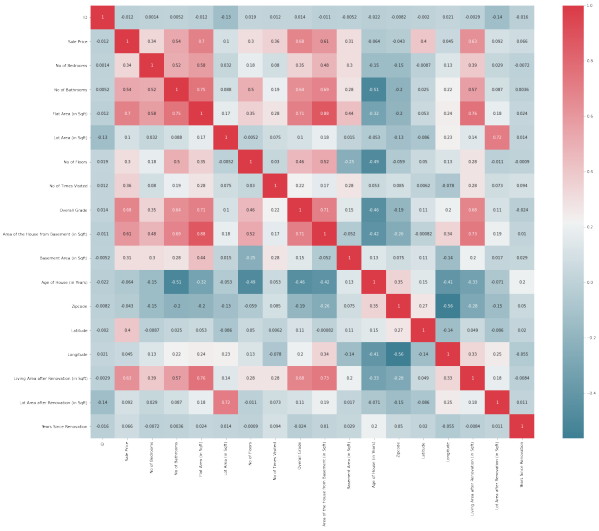

Figure 2. Correlation Map

3.1.3 Categorical Variables

- Find the frequency distribution of unique values of a categorical variable.
- Visualize the relationship between a numerical target variable and a categorical variable with the help of a bar chart.
- Identify all the categorical variables :
  - Waterfront View
  - Condition of the House
  - Zipcode
  - Ever Renovated
- The questions one needs to answer while exploring categorical variables
  - What information does this variable contain
  - Do the few sample values that we are able ‘eyeball’
  - How many and what unique values , i.e, level, does this categorical variable have and what is the frequency distribution of those unique values
  - How is dependent/target variable correlated with this variable.
- The method of finding out whether a Categorical variable is having an impact on the numeric variable is known as **ANOVA**
- **ANOVA**- it checks if the means of the target variables across different levels or unique values of a categorical variable are equal or not.

- **Binning and Creation of dummy variables:**
  - Transforming a categorical variable into a set of numerical or boolean variables called dummy variables, each of which has values of 0 or 1.
  - If Categorical variables has 'n' levels → n-1 dummies are required

  Why we are doing this :
  - Regression modeling requires all the independent variables to be numerical variables Get_dummies of pandas library creates the dummy variable and drop the categorical variable. Binning - if number of levels > 20 , then bin these levels into fewer groups before creating the dummy variables.
  - Used for both numerical and categorical variables.

3.2 Feature Scaling

Feature Scaling is a must for those algorithms where some measure of distance between data points is involved like Logistic Regression, Linear Regression, K Nearest Neighbors, Principal Component Analysis, etc.

Different Techniques of Feature Scaling:

- Standardisation : It rescales the feature values so that they have the properties of a Standard Normal Distribution with mean as 0 and the standard deviation of 1.
- Min Max Scaling: One of the simplest methods for scaling.The value range for the transformed variables lies between [0, 1]
- Normalization: The range is fixed from -1 to 1. Also called mean normalization
- The preprocessing library of sklearn can be used to do the standardisation.
- Standardisation method should be used for scaling the feature variables when a linear regression model is built.
- A linear regression model assumes the input variables to be normally distributed.
-

3.3 Feature Engineering

we conducted set of feature engineering step such as:

- The technique of generating new features using the existing features is called feature engineering.

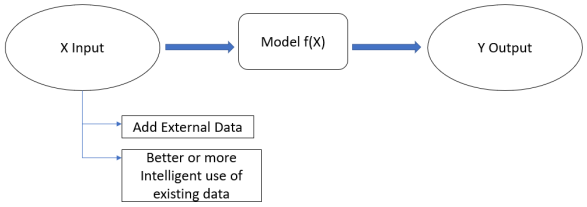

Figure 3. Flowchart: Feature Engineering

- The science of extracting more information from existing data

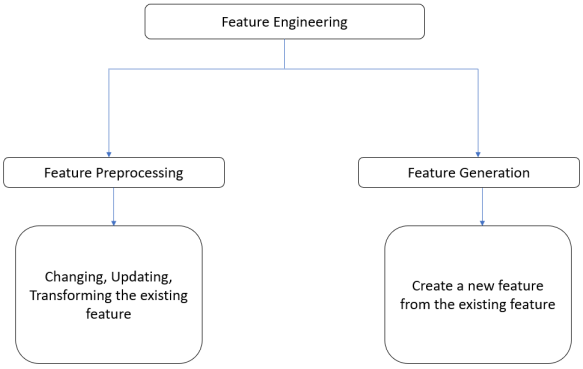

Figure 4. Types of Feature Engineering

3.4 Linear Regression

If model have the linear relationship between the dependent variables.For finding a relationship between two continuous variables, Linear regression is useful. One variable is predictor or independent, and the other variable is variable response or dependent. Parameters :

$$Y = mX + C \tag{1}$$

Where:

- Y → Target Variable
- X → Independent Variable
- m → Slope
- C → Intercept

m, C are the parameters

Assumptions of linear regression

1. The target and the independent variable should have a linear relation between them. It need not be a perfect linear relation (i.e correlation = 1). But should follow a fundamental property of a linear relation.

- Fundamental Property of Linear Relation

  Y changes with X and for every change of AX, the quantum of change in Δ i.e. A should be similar irrespective of value of X. At X = X, a change of ΔY causes change in value of Y, then at X = X also a change of ΔX, should cause similar change as ΔY.

  Make the non-linear relationship linear by using variable transformation operations such as x^2, x or log(x)

2. Constant Variance of Error (Homoscedastic)
3. Normal distribution of Errors
4. No correlation between Error Terms
5. No multicollinearity

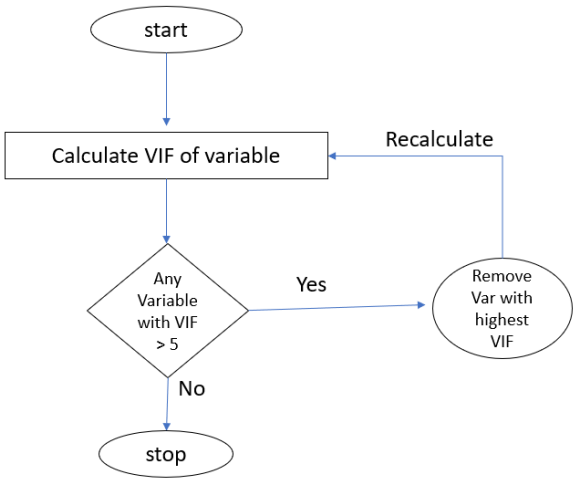

Figure 5. Flowchart: Treating Multicollinearity using VIF

Overview In this implementation, we will be looking at the following steps:

1. Importing Libraries and Dataset :
   Python libraries like :
   - Numpy
   - Pandas
   - Matplotlib
   - Sklearn
   - seaborn
   - statsmodel
2. Scaling the dataset
   - The preprocessing library of sklearn can be used to do the standardisation.
3. Checking Multicollinearity and removing it
   - Removing only one variable from the pair of correlated independent variables is enough.We will use variance inflation factor (VIF) to treat Multicollinearity.

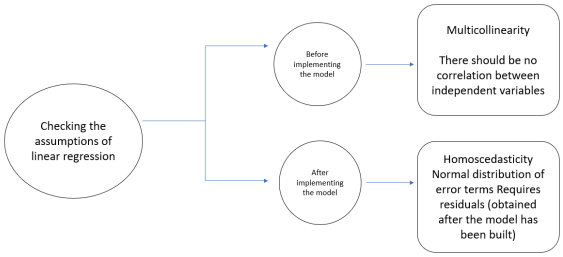

Figure 6. assumption of linear Regression

4. Creating test and training partitions
   Data-
   - Training Dataset : To educate or train our models
   - Test Dataset : To examine model performance

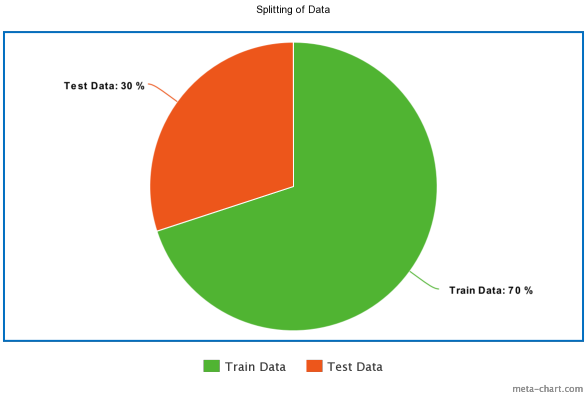

Figure 7. Pie chart for splitting of Dataset

5. Implementing the Linear Regression model
   - Using the sklearn library
   - This library allows us to implement the linear regression model without having to write the code for gradient descent
   - Lr.fit → Implements Gradient descent and the complete procedure over the training data

6. Generating predictions over the test set
7. Evaluating the model
8. Preparing the Residual plot.
9. Checking the Assumptions of Linear Regression.
    - Making Residual Plots to verify assumptions
    - Earlier, no independent variable was used while plotting the residual plot of the mean regression , but in case of linear regression we need to make it.

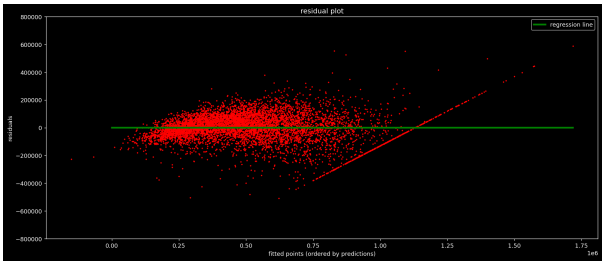

Figure 8. Residual of linear regression model

10. Visualising the Coefficient plot

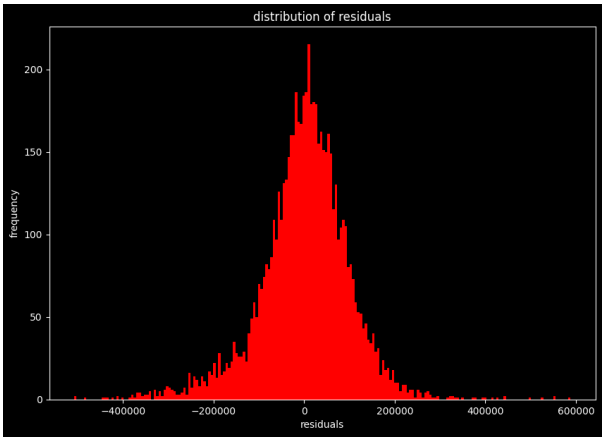

Figure 9. distribution of residuals

3.5 Gradient Descent

An optimization algorithm that works iteratively and aims to find the minimum value of a convex function with respect to a set of parameters.

Steps:

1. Random initialization
2. Take the minimum cost Iterativelyialization
3. Generating Prediction
4. Calculating cost/error
5. Updating parameters
6. Repeat form step 2
7. Take the minimum cost Iteratively

Gradient boosting is a machine learning technique for regression and classification issues that generates a predictive model in the form of a set of weak predictive models, typically decision trees.

3.6 Model Evaluation Metrices

1. MAE - mean absolute error:
    - Take the absolute value of each individual error term, sum it up for all the data points and then take a mean.
    - The mean absolute error tells us how far, on an average, the actual point is expected to lie from the predicted point.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

Figure 10. Mean Average Error

2. MSE - mean squared error
    - It turn all the term into positive quantity
    - It incurs extra penalty for the larger differences between the actual and the predicted values
    - It fails the expected distance between the actual points and the predictions.
    - Square the individual errors, sum them up, and then take a mean of it.

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

Figure 11. Mean Squared wrror

3. RMSE- Root mean squared error

- The large errors are being penalised, yet the scale of the error is closer to that of mean absolute error..

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

Figure 12. Root mean squared error

4. R^2 - R squared

- The large errors are being penalised, yet the scale of the error is closer to that of mean absolute error.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Figure 13. R squared

5. Adjusted R^2 - Adjusted R squared

- It penalizes the result for adding variables which do not improve your existing model.

$$Adjusted\ R^2 = 1 - (1 - R^2) * \frac{n-1}{n-k-1}$$

Figure 14. Adjusted R squared

4 Result and Conclusion

In this paper we explored linear Regression models to predict the Sale Price of a house based on its characteristics given.
We have our model coefficients:
array([-3928.66247639, 12028.44560689, 14967.00497585, 2697.55278605, 27220.31313417, 59965.44665815, 80697.80906997, 27729.56715434, 27873.90231343, 21397.40341959, -23854.32640243, 17943.26729788, -2896.98542901, -10179.085198 , 14239.3533334 , 5095.97603572, -2296.64888137, 14594.33847962, 10761.77007875, 12165.83372082, 33842.29544383, 63269.82875283, 81086.08553213, 50718.63947886, 73274.09568028, 40153.03595158, 67405.70271285, 22113.74944051])
and linear regression model score as:
0.8461987715586199
we got an accuracy of more than 84 percent that is respectable score to go with.
Predictions: [550000.54013871 645339.25748139 612230.85487286 ... 1026223.16735842 944048.4100588 1204608.7500112]

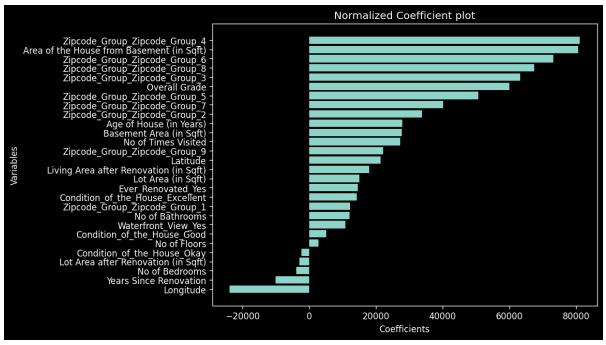

Figure 15. Model coefficient Plot

5 Acknowledgement

We would like to Thank prof. Ritesh Kumar Mishra (Prof. NIT Patna) who had been constantly supporting and mentorinig us and also helped shape our report which helped in giving us much deeper insights into the course which eventually helped us in the completion of this report.

References

[1] https://www.irjet.net/archives/V6/i4/IRJET-V6I4677.pdf/

[2] https*//towardsdatascience.com/house-price-prediction-with-zillow-economics-dataset-*18709*abff*896

[3] trainings.internshala.com/progress/home/machine-learning

[4] https://www.k2analytics.co.in/multiple-linear-regression-adjusted-r-squared/

[5] https://www.datatechnotes.com/2019/02/regression-model-accuracy-mae-mse-rmse.html