1. Introduction

1.1 Background

There are many people who are planning to buy their dream house. These people have preferences how their house will look like and what it contains. So, who is not having good knowledge of houses faces difficulty to find good houses with good price Therefore, through this project the task for selecting a house made easy.

1.2 Problem

The problem is to predict price to select a good house. For example, the person may select a house that doesn’t worth it, then we can select house based on some attributes according to our preferences and predict its price. Considering such challenges there is a requirement for a model which can provide with the right option for selecting a house with good rate.

2. Data

The data used in this project is taken from Kaggle site which is a train and test data used for our model building to predict price of the house.

2.1. Data Understanding

There are total 80 variables in train data in which 79 are features or independent variables and one SalePrice is a target or dependent variable. Variables and their data type is shown as:

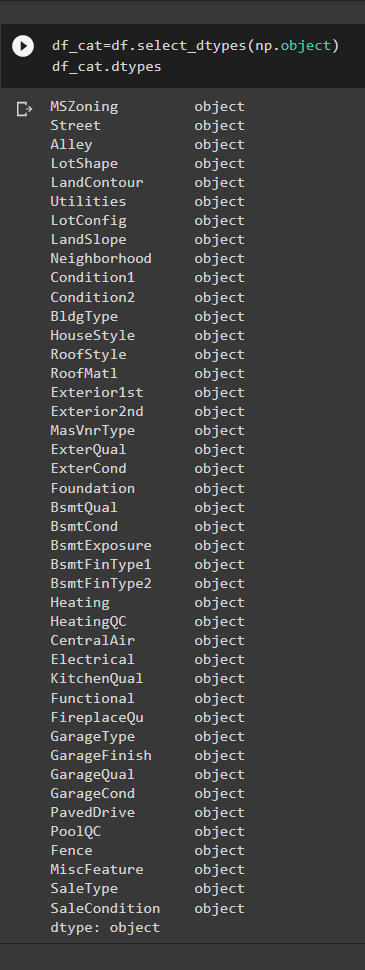
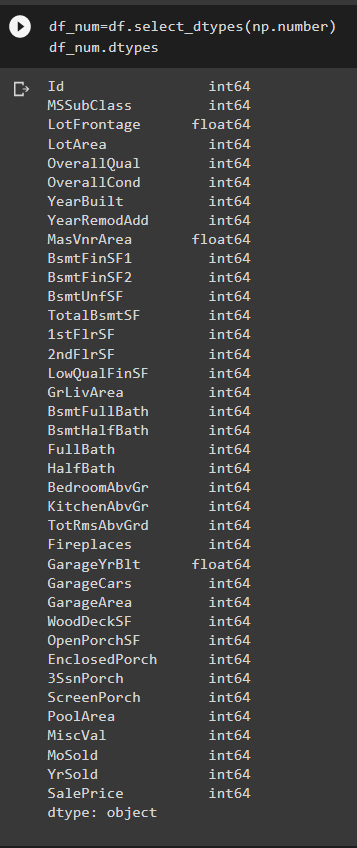


Fig: Numerical (left) & Categorical(right) variable

2.2. Exploratory Data Analysis

In this part the proper analysis of the type and origin of data is performed. The relationship between variable is explored. The relationship between the attributes is examined via the frequency. We see the summary of data. And dropped some variable which are not appropriate.

2.3. Data Preparation

- Here we are dealing with the null or missing values in the data, If the missing values is more than 80% we dropped that variable and if it is less than 80% we will do the imputation on them. We can see this through heatmap which is shown as:

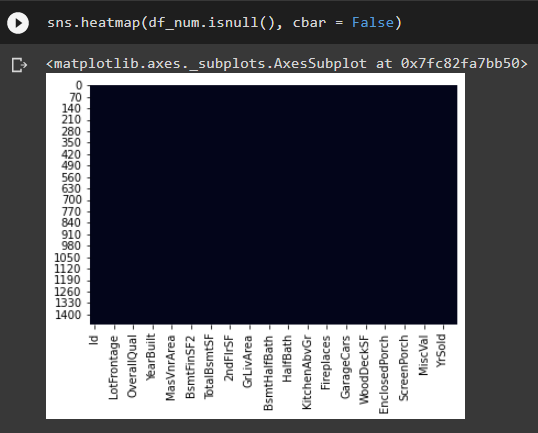
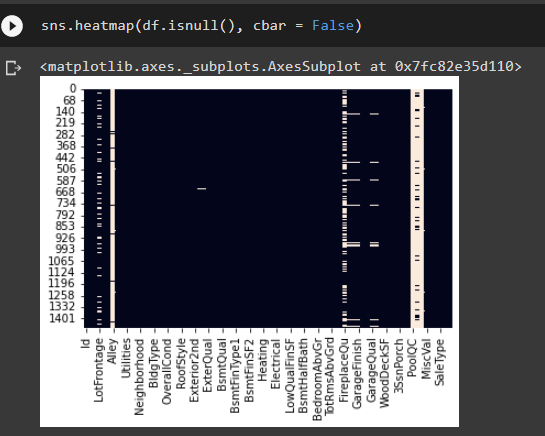


Fig: Before removing null(left) & After removing null(right)

- After that we deal with the outliers and remove the extreme values through inter quartile range.

- Then, we do the preparation of data for modelling in which we scale the numerical data and target variable to bring it in a scalable range and do the dummy encoding on the categorical variable and split the data in 80:20 ratio.

3. Model

Here we use basic linear regression algorithm on our train data to predict price of house.

- First we built a simple linear regression model,

- Second we built model on selecting features based on p-value,

- Third we built model using stochastic gradient descent

- And finally we built model on features which is selected by recursive feature selection technique.

4. Result

Now, we predict the price of test data and calculate their rmse score so decide which is good model as for now. Here is the table of model with their rmse score:

|  |  |
| --- | --- |
| Model | RMSE score |
| Simple Linear Regression | 12.51 |
| Linear Regression using p value | 12.25 |
| Linear Regression using stochastic gradient descent | 11.77 |
| Linear Regression using recursive feature selection technique | 11.65 |

5. Conclusion

As we can see from above table that using recursive feature selection technique we get some good results. Further we can improve our model by applying some regularization technique like ridge, lasso, elastic net. We can also improve this by using some more feature engineering to select appropriate features and finally we can also use some advance regression techniques like random forest, gradient boost, XG boost technique to improve or model.