Housing Price Prediction

Submitted By :

MANISHA KUMARI

From my understanding the problem is about making prediction of house prices for the company which is about to buy that property & to help make decision that whether it would be profitable or not to buy it on the basis of other variables given so that we could make this venture more profitable for the company.

## Importing Libraries for the task

Numpy, pandas, for imported data

## Importing Dataset using libraries

Imported the datasets using pandas library in jupyter notebook.

## Checking Dimension of dataset

By checking dimension of train dataset we get to know that it contains 1161 rows & 81 columns & for test dataset we get to know that it contains 292 rows & 80 columns.

## Checking Description of dataset

From description we find mean, min value, max value, etc of every column which contains continuous data in them

## Checking for presence of null values in dataset

We checked for the presence of null values in every column of dataset by doing it repeatedly for every column as doing it repeatedly for every column & null values were present. Then we filled them randomly to avoid data bias.

## Identifying Target variable

By looking at dataset we identified target variable which is named SalePrice.

## Performing EDA on whole Train dataset Id

From visualization we conclude that could not determine the skewness of data with data having highest density at 600 & we could not determine the correlation of the data with target column (SalePrice) as data is scattered all over the place.

## Plotting Heatmap

We plotted heatmap of train dataset to check correlation of every column with target column & we found that many columns having high positive correlation with target column. Also we dropped following ( Utilities, SaleCondition, Alley, GarageQual) columns to reduce multicollinearity.

## Preparing Test dataset

After plotting heatmap for train dataset, we filled the null values in test dataset & changed the data types of every column to int or float data type present in test dataset. Also target column was not present in test dataset.

## Scaling the data

After preparing both train & test dataset we scaled them to remove outliers from them.

## Applying GridSearchCV

We chose 4 regression models (Linear Regression, Decision Tree Regression, Random Forest Regression, Bagging Regression) to check performance of them on the dataset to find out best suited model for the dataset. Then one by one we applied GridSearchCV on each model to find best hyper parameter tuning while model working on dataset. After applying GridSearchCV on every model we concluded that Random Forest Regression model is giving best train accuracy which was 0.9110 with hyper parameter combination (max\_depth : 5, min\_samples\_leaf : 1, min\_samples\_split : 2)

## Applying Model on dataset

Chose Random Forest Regression Model & made it ready by tuning hyper parameter. After tuning hyper parameter we applied model on dataset to make prediction. After making prediction we made a dataframe which contained predicted value as test dataset does not have the target column. So we could’nt check the performance of the model.

## Saving Model

After making prediction we saved the model using pickle library.

# CONCLUSION

In the end we conclude that model is quite accurate in predicting the price of houses & price is getting affected positively because of the following variables (OverallQual, YearBuilt, YearRemodAdd, ExterQual, GrLivArea).