

Housing Project

Submitted by:

KARISHMA YADAV.

**ACKNOWLEDGMENT**

I have taken efforts in this project. A housing project that effectively manages and handles all the functioning of a cooperative housing society. The software system can store the data of various flat owners and their family members along with their images.

The Initiation Process consists of those processes performed to define a new project or a new phase of an existing project by obtaining agreement to start the project or phase.

**INTRODUCTION**

Houses are one of the necessary need 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 some of the machine learning techniques used for achieving the business goals for housing companies. Our problem is related to one such housing company.

1. Regression analysis: Regression analysis techniques aim mainly to investigate and estimate the relationships among a set of features. Regression includes many models for analysing the relation between one target/response variable and a set of independent variables. Logistic Regression (LR) is the appropriate regression analysis model to use when the dependent variable is binary. LR is a predictive analysis used to explain the relationship between a dependent binary variable and a set of independent variables.
2. Decision Tree: Decision Tree (DT) is a model that generates a tree-like structure that represents set of decisions. DT returns the probability scores of class membership. DT is composed of: a) internal Nodes: each node refers to a single variable/feature and represents a test point at feature level; b) branches, which represent the outcome of the test and are represented by lines that finally lead to c) leaf Nodes which represent the class labels. That is how decision rules are established and used to classify new instances. DT is a flexible model that supports both categorical and continuous data. Due to their flexibility they gained popularity and became one of the most commonly used models for prediction.
3. Random Forest Random forests (RF) are an ensemble learning technique that can support classification and regression. It extends the basic idea of single classification tree by growing many classification trees in the training phase. To classify an instance, each tree in the forest generates its response (vote for a class), the model choses the class that has receive the most votes over all the trees in the forest. One major advantage of RF over traditional decision trees is the protection against over fitting which makes the model able to deliver a high performance.
4. KNN

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.

The first step before applying the selected analytical models on the dataset, explanatory data analysis for more insights into dataset was performed. Based on the observations data was pre-processed to be more suitable for analysis

**Analytical Problem Framing**

Data pre-processing:

Pre-processing includes three steps:

a) Data transformation,

b) Data cleaning and

c) Feature selection.

Data Transformation :

Two of the explanatory variables were transformed from binominal form (yes/no) into binary form (1/0) to be more suitable for the selected models.

Data cleaning :

This stage includes missing data handling/imputation: Some of the selected algorithms cannot handle missing data such as SVM. That’s why missing value can be replaced by mean, median or zero. However, missing data replacement by statistically computed value (imputation) is a better option. The used dataset included missing values in some the numerical variables. Numerical data were replaced using random forest imputation technique And binary values were imputed using the techniques.

Feature selection :

Before model training, feature selection is one of the most important factors that can affect the performance of models. In this study, the importance of the used variables was measured to identify and rank explanatory variables influence on the target/response. This allows dimensionality reduction by removing variables/predictors with low influence on the target. Random forest technique can be used for feature selection using mean decrease accuracy. Mean decrease measures the impact of each feature on model accuracy. The model permutes values of each feature and evaluates model accuracy change. Only features having higher impact on accuracy are considered important.

* Hardware and Software Requirements and Tools Used

RAM: 8GB

ROM: I3 processor

SOFTWARE: Python 3.9.6

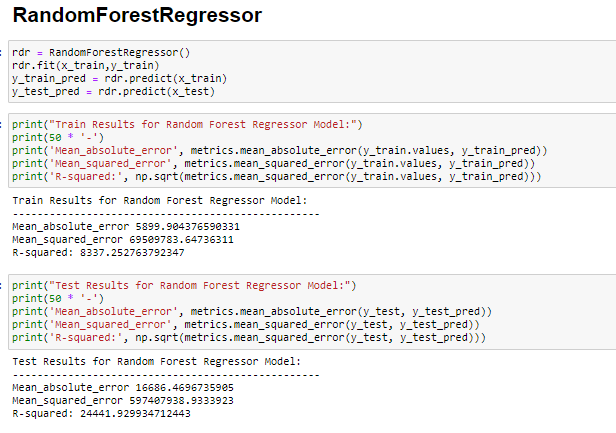
LIBRARIES: Numpy, Pandas, Seaborn, Sklearn, Scipy .

Tools: Jupiter notebook, Matplotlib, Scikit-learn, Excel.

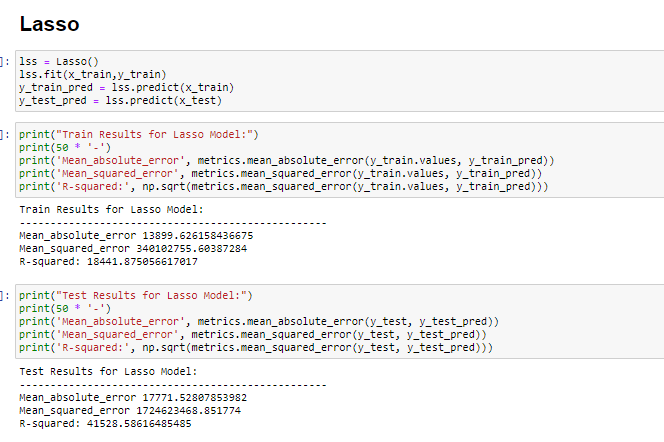
* Testing of Identified Approaches (Algorithms)

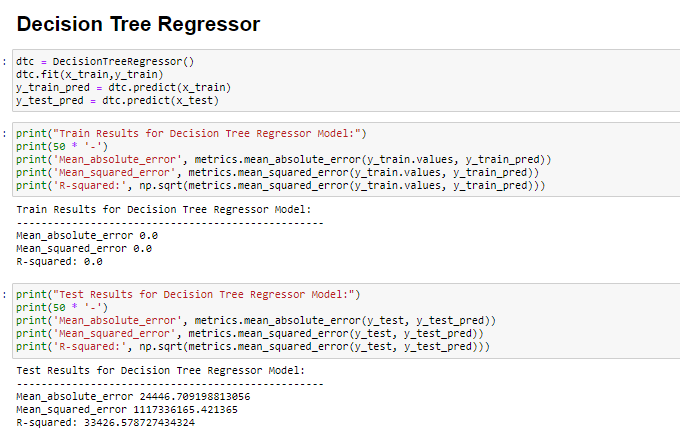
1. K-Neighbors Regressor
2. Decision Tree Regressor
3. Random Forest Regressor
4. Ridge
5. Lasso

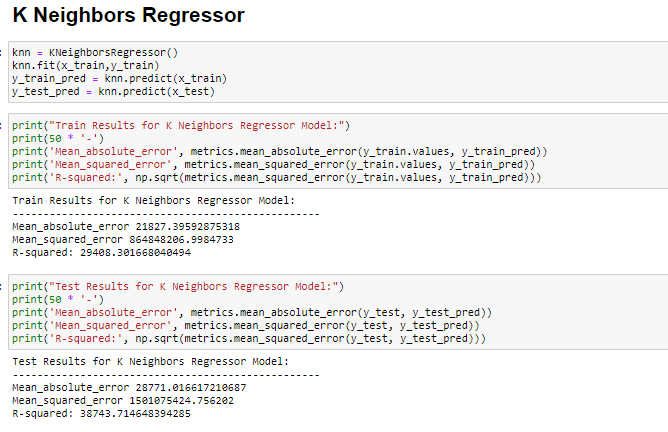
* Run and Evaluate selected models
* MODELS





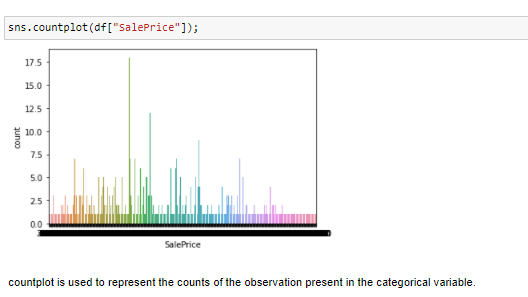




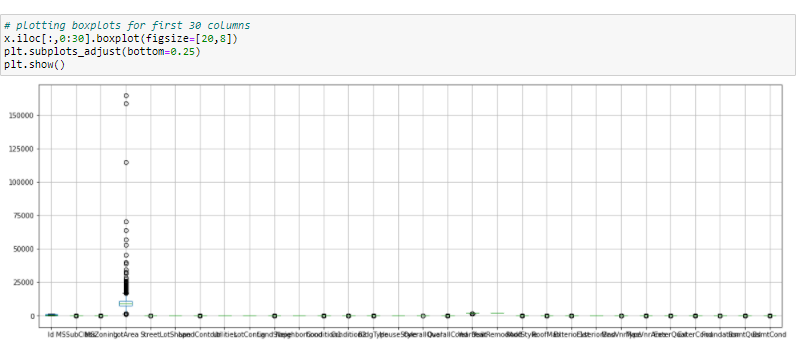


* Visualizations

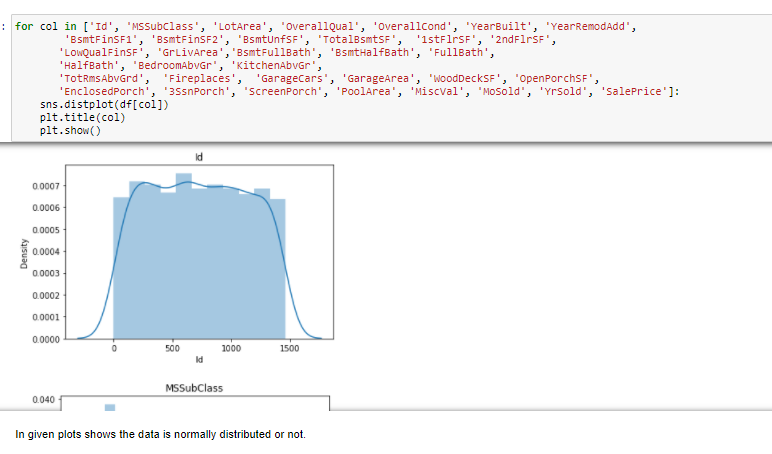
1. Count plot
2. Boxplot
3. Distplot
4. Scatterplot
5. Heatmap
6. Count plot



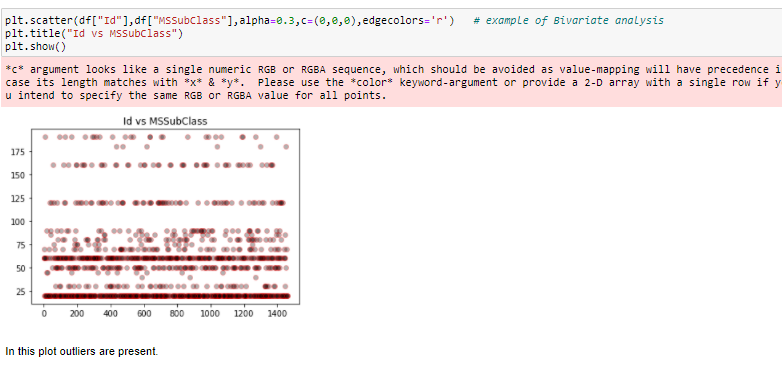
1. Boxplot



1. Distplot



1. Scatterplot



1. Heatmap



**CONCLUSION**

In these project, Decision Tree Regressor, Random Forest Regressor, K-Neighbors Regressor, Ridge, Lasso these models are used.

Random Forest Regressor is r2 score ,Mean absolute error ,Mean\_squared\_error and R-squared is good than other models, Hence Random Forest Regressor is performing is good.

Random Forest Regressor is best model for this dataset.

