**HOUSE PRICE PREDICTION USING VARIOUS REGRESSION :**

**A COMPARATIVE STUDY**

**ABSTRACT**

House price forecasting is an important topic of real estate. The literature attempts to derive useful knowledge from historical data of property markets. Machine learning techniques are applied to analyze historical property transactions in India to discover useful models for house buyers and sellers. Revealed is the high discrepancy between house prices in the most expensive and most affordable suburbs in the city of Bangalore. Moreover, experiments demonstrate that the Linear Regression that is based on mean squared error measurement is a competitive approach.

# **INTRODUCTION**

**1.1 Motivation**

We are highly interested in anything related to Machine Learning, the independent project provided us with the opportunity to study and reaffirm our passion for this subject. The capacity to generate guesses, forecasts, and offer machines the ability to learn on their own is both powerful and infinite in terms of application possibilities. Machine Learning may be applied in finance, medicine, and virtually any other field. That is why we opted to base our idea on Machine Learning

**1.2 Objective**

As a first project, we intended to make it as instructional as possible by tackling each stage of the machine learning process and attempting to comprehend it well. We have picked Bangalore Real Estate Prediction as a method, which is known as a "toy issue," identifying problems that are not of immediate scientific relevance but are helpful to demonstrate and practice. The objective was to forecast the price of a specific apartment based on market pricing while accounting for various "features" that would be established in the following sections.

**METHODOLOGY**

**2.1 Data Collection**

The statistics were gathered from Bangalore home prices. The information includes many variables such as area type, availability, location, BHK, society, total square feet, bathrooms, and balconies

2.2 **Data Exploration**

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, Python. Before it can conduct analysis on data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working.

We divided the data 9:1 for Training and Testing purpose respectively.

**2.3 Data Selection**

Data selection is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data. Data selection precedes the actual practice of data collection. This definition distinguishes data selection from selective data reporting (selectively excluding data that is not supportive of a research hypothesis) and interactive/active data selection (using collected data for monitoring activities/events or conducting secondary data analyses). The process of selecting suitable data for a research project can impact data integrity.

The primary objective of data selection is the determination of appropriate data type, source, and instrument(s) that allow investigators to adequately answer research questions. This determination is often discipline-specific and is primarily driven by the nature of the investigation, existing literature, and accessibility to necessary data sources.

**2.4 Linear Regression**

Linear regression is a supervised learning technique. It is responsible for predicting the value of a dependent variable (Y) based on a given independent variable (X). It is the connection between the input (X) and the output (Y). It is one of the most well-known and well-understood machine learning algorithms. Simple linear regression, ordinary least squares, Gradient Descent, and Regularization are the linear regression models.

**2.5 Lasso Regression:**

Lasso regression is another regularization technique to reduce the complexity of the model. It stands for Least Absolute and Selection Operator. It is similar to the Ridge Regression except that the penalty term contains only the absolute weights instead of a square of weights.

**2.6 Decision Tree Regression**

It is an object that trains a tree-structured model to predict data in the future in order to provide meaningful continuous output. The core principles of decision trees, Maximizing Information Gain, Classification trees, and Regression trees are the processes involved in decision tree regression. The essential notion of decision trees is that they are built via recursive partitioning. Each node can be divided into child nodes, beginning with the root node, which is known as the parent node. These nodes have the potential to become the parent nodes of their resulting offspring nodes. The nodes at the informative features are specified as the maximizing information gain, to establish an objective function that is to optimize the tree learning method.

**PROJECT**

**3.1 Problem Statement**

Create a model to estimate the price of houses in Bengaluru.

**3.2 Data**

The data is the most important aspect of a machine learning assignment, to which special attention should be paid. Indeed, the data will heavily affect the findings depending on where we found them, how they are presented, if they are consistent, if there is an outlier, and so on. Many questions must be addressed at this stage to ensure that the learning algorithm is efficient and correct.

To obtain, clean, and convert the data, many sub steps are required. We will go through these steps to understand how they've been used in my project and why they're helpful for the machine learning section.

**3.3 Dataset**

<https://github.com/amrutkar20/banglore_home_prices_prediction_model/blob/main/Bengaluru_House_Data.csv>

**3.4 Model**

<https://github.com/amrutkar20/banglore_home_prices_prediction_model/blob/main/banglore_home_prices_prediction_model.ipynb>

**3.5 Project Architecture**

Diagram

Description automatically generated

**3.6 Data Science**

The first stage is standard data science work, in which we take a data set named ‘Bengaluru House pricing data' from Kaggle. We will do significant data cleaning on it to guarantee that it provides reliable predictions throughout prediction. This Jupiter notebook, ‘Bangalore-House Price-Prediction-Model.ipynb,' is where we do all of our data science work. Because the Jupiter notebook is self-explanatory, we will only touch on the principles that we have implemented briefly. In terms of data cleansing, our dataset needs a significant amount of effort. In fact, 70% of the notebook is dedicated to data cleaning, in which we eliminate empty rows and remove superfluous columns that will not aid in prediction.

The process of obtaining valuable and significant information from a dataset that will contribute the most to a successful prediction is the next stage.

The final stage is to deal with outliers. Outliers are abnormalities that do massive damage to data and prediction. There is a lot to comprehend conceptually from the dataset in order to discover and eliminate these outliers.

Finally, the original dataset of over 13000 rows and 9 columns is reduced to about 7000 rows and 5 columns.

**3.7** **Machine Learning**

The resulting data is fed into a machine learning model. To find the optimal procedure and parameters for the model, we will mostly employ K-fold Cross-Validation and the GridSearchCV approach.

It turns out that the linear regression model produces the best results for our data, with a score of more than 80%, which is not terrible.

Now, we need to export our model as a pickle file (Bengaluru\_House\_Data.pickle), which transforms Python objects into a character stream. Also, in order to interact with the locations(columns) from the frontend, we must export them into a JSON (columns.json) file.

**EXPERIMENTAL SETUP**

**4.1 Steps to Create Model**

1. Import Libraries

2. Load Dataset

3. Exploratory Data Analysis

4. Data Cleaning

5. Feature Engineering

6. Dimensionality Reductions

7. Outlier Removal using Business Logic

8. Outlier Removal using Standard Deviation & Mean

9. Data Visualization

10. Building a Model

11. Test the Model for few properties

12. Export the tested model to a pickle file

**4.2 Tools used**

1. Anaconda

2. Jupiter Notebook

3. Google Collaboratory

**4.3 Technologies used**

1. Python

# **CONCLUSION**

From the above made model, we can conclude that the Linear Regression algorithm performs better in performing prediction of the real-estate price than Lasso Regressor and Decision Tree Regressor with an accuracy of 82% and in some cases, it increased up to 89%.

Hence a high accuracy model was built with basic data cleaning steps and machine learning algorithms.

**REFERENCE**

[https://www.javatpoint.com/machine- learning](https://www.javatpoint.com/machine-%20%20%20%20learning)<https://www.kaggle.com/datasets/amitabhajoy/bengaluru-house-price-data>