Predicting House Prices Using Machine Learning....

Abstract:
Predictive models for determining the sale price of houses in cities like Bangalore is still remaining as more challenging and tricky task. The sale price of properties in cities like Bangalore depends on a number of interdependent factors. Key factors that might affect the price include area of the property, location of the property and its amenities. In this research work, an analytical study has been carried out by considering the data set that remains open to the public by illustrating the available housing properties in machine hackathon platform. The data set has nine features. In this study, an attempt has been made to construct a predictive model for evaluating the price based on the factors that affect the price. Modeling explorations apply some regression techniques such as multiple linear regression (Least Squares), Lasso and Ridge regression models, support vector regression, and boosting algorithms such as Extreme Gradient Boost Regression (XG Boost).

I. Introduction

Modeling uses machine learning algorithms, where machine learns from the data and uses them to predict a new data. The most frequently used model for predictive analysis is regression. As we know, the proposed model for accurately predicting future outcomes has applications in economics, business, banking sector, healthcare industry, e-commerce, entertainment, sports etc. One such method used to forecast house prices are based on multiple factors . In metropolitan cities like Bangalore, the prospective home buyer considers several factors such as location, size of the land, proximity to parks, schools, hospitals, power generation facilities, and most importantly the house price. Multiple linear regressions is one of the statistical techniques for assessing the relationship between the (dependent) target variable and several independent variables. Regression techniques are widely used to build a model based on several factors to predict price. In this study, we have made an attempt to build house price prediction regression model for data set that remains accessible to the public in Machine hackathon platform. We have considered five prediction models, they are ordinary least squares model, Lasso and Ridge regression models, SVR model, and XG Boost regression model. A comparative study was carried out with evaluation metrics as well**.**

Evaluate resources
Building the plan
Identification of target market and properties
Executing the acquisitions

Data preprocessing:

Data preprocessing is a predominant step in machine learning to yield highly accurate and insightful results. Greater the quality of data, greater is the reliance on the produced results. **Incomplete, noisy, and inconsistent data** are the properties of large real-world datasets. Data preprocessing helps in increasing the quality of data by filling in missing incomplete data, smoothing noise and resolving inconsistencies.

- **Incomplete data** can occur for a number of reasons. Attributes of interest may not always be available, such as customer information for sales transaction data. Relevant data may not be recorded due to a misunderstanding, or because of equipment malfunctions.
- There are many possible reasons for **noisy data** (having incorrect attribute values). The data collection instruments used may be faulty. There may have been human or computer errors occurring at data entry. Errors in data transmission can also occur. Incorrect data may also result from inconsistencies in naming conventions or data codes used, or inconsistent formats for input fields, such as date.

There are a number of data preprocessing techniques available such as,

1. **Data Cleaning**
2. **Data Integration**
3. **Data Transformation**
4. **Data Reduction**

Data Acquisition:

This is a **House Price Prediction Competition**. The objective of the project is to perform data visualization techniques to understand the insight of the data. Machine learning often required to getting the understanding of the data and its insights. This project aims apply various tools to get a visual understanding of the data and clean it to make it ready to apply machine learning and deep learning models on it.

Problem Statement :

Housing prices are an important reflection of the economy, and housing price ranges are of great interest for both buyers and sellers. Ask a home buyer to describe their dream house, and they probably won’t begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition’s data-set proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

Approach taken:

We’ll use the Random Forest regression algorithm to predict the price of the houses. In this article, we’ll consider machine learning algorithms as a black box that fits the data.

This article focuses more on the machine learning pipeline. For more information on the Random Forest algorithm, I suggest looking into this video. We’ll begin by loading the data. Since we’re using an inbuilt dataset, we’ll be calling the load Boston function from the sklearn.datasets module. We load the data into the data variable.

Once the data is loaded, we separate the data and target attributes of the data variable. We store them in variables data and target, respectively.

Conclusion

We have gone through how to implement the entire machine learning pipeline, and we have an intuitive understanding of machine learning algorithms. The larger the dataset gets, the more complex each of the mentioned steps gets. Therefore, using this as a base will help while you build your knowledge of machine learning pipelines.