House Price Prediction using Machine Learning

Submitted to – Mr. Anurag Goel

Prajjwal Chittori[*], Pranav Roy[**]

[*] Department of Computer Science & Engineering, Delhi Technological University
[**] Department of Computer Science & Engineering, Delhi Technological University

Abstract - This project tries to compare the performances of various machine learning regression algorithms and classification techniques. This project indulges primarily in Multiple Linear regression and Keras Regression. The dataset of the project is acquired from Kaggle of house prices in Bangalore (https://www.kaggle.com/amitabhajoy/bengaluru-house-price-data).
There are 10 total attributes in the dataset, four of which we derived from current columns. The models will be using 10 attributes.

Keywords - Multiple linear regression, Keras Regression, Artificial Neural Network, Machine Learning, House Price Prediction.

INTRODUCTION

Machine Learning is a subfield of AI which is used to extract useful information from data.
Machine learning techniques are suitable for big data because manually processing massive amounts of data would be difficult without the assistance of computers. In computer science, machine learning tries to solve problems algorithmically rather than mathematically. As a result, it is focused on the development of algorithms that enable the computer to learn.
Since several regression algorithms rely on an unknown number of attributes to produce the value to be expected, the output would be calculated by predicting house prices. The cost of a home is determined by its specifications. Houses have a variety of features that may or may not have the same cost depending on where they are located. For example, a large house in a desirable rich neighborhood can command a higher price than one in a poor neighborhood.

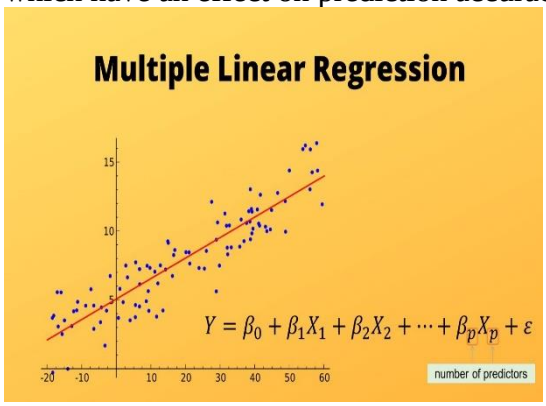
AIM

The aim of this study is to examine the accuracy of predicting house prices using multiple linear, and Artificial Neural Network regression algorithms (ANN). As a result, the aim of this research is to gain a better understanding of regression methods in machine learning.

Background

I. Multiple Linear Regression

MLR (Multiple Linear Regression) is a supervised technique for estimating the relationship between multiple independent variables and one dependent variable. Identifying the association and its cause-effect relationship aids in the prediction process. When estimating these relationships, the model's prediction accuracy is critical; the model's complexity is of greater importance. Multiple Linear Regression, on the other hand, is vulnerable to a number of issues, including multicollinearity, noise, and overfitting, all of which have an effect on prediction accuracy.

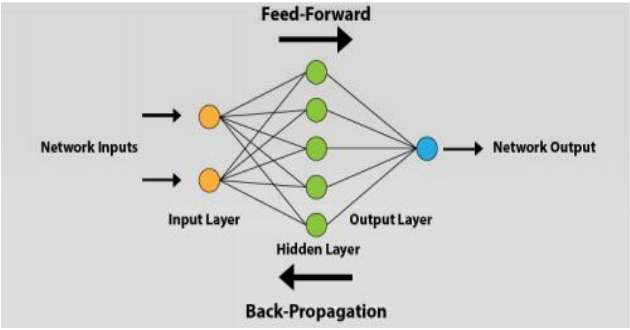


II. Keras Regression

Keras is a programming interface for high-level neural networks. The model was created by Google and runs on top of TensorFlow.
At the moment, PyTorch, created by Facebook, is Keras' main competitor. Though PyTorch has a larger community, it is an especially verbose language, and I prefer Keras for its greater simplicity and ease of use when creating and deploying models.

III. ANN

An artificial neural network (ANN) is a computer program that attempts to mimic the functions of a biological brain.
Neurons are a set of interconnected elements or nodes that make up an ANN. Neurons function as channels, taking in information, processing it, and passing it on to other neurons for further processing.
Layers handle this exchange, or the process of transmitting data between neurons. The input layer, one or more hidden layers, and the output layer are the three layers that make up a layer. Each layer contains a collection of neurons that take in data, process it, and then transfer the results to the next layer's neurons. This process is repeated until you enter the output layer.
As a result, ANN attempts to simulate the brain's ability to learn a pattern in a given dataset and predict its performance, regardless of whether the predicted data was presented during the learning process.



Method

Theoretical analysis and realistic implementation of regression algorithms have been used to organise this study.

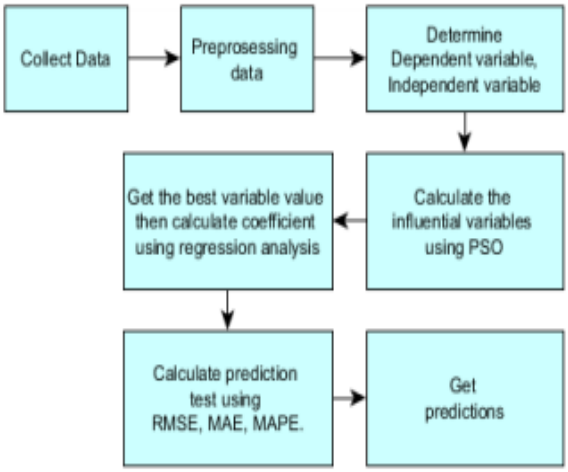
I. Literature Study

Instead of textbooks and chapters of books, the majority of the literature review is focused on papers with full text online, open access articles, and peer-reviewed publication.The aim of this literature review is to build a solid foundation for regression techniques, regularisation, and artificial neural networks in machine learning, as well as how they can be applied precisely to house price prediction.

II. Experiment

The experiment is carried out to pre-process the data and assess the models' prediction accuracy. To obtain the prediction results, the experiment must go through several steps. There are

Pre-processing: both datasets will be reviewed and pre-processed using the methods. These methods handle data in a variety of ways. As a result, the preprocessing is performed in several iterations, with the accuracy being tested each time with the used combination.
Data splitting: splitting the dataset into two sections is necessary in order to train the model with one and evaluate it with the other. 75 percent of the dataset will be used for training and 25% for research.
The accuracy of both datasets will be tested by comparing the real prices on the test dataset to the prices expected by the model, as well as calculating the R2 and RMSE rate while training the model.
Performance: in addition to the measurement parameters, the time taken to train the model will be calculated in order to demonstrate how the algorithm varies over time.
Correlation: The Pearson Coefficient Correlation will be used to determine if the available features have a negative, positive, or zero correlation with the house price.



Regression Analysis

Hedonic pricing was used as the prediction model in this study, which is a suitable regression model with the standard formula shown in . price is the dependent variable, and year, building area, land area, NJOP land price (IDR/m 2), NJOP building price (IDR/m 2), distance to city centre, number of campuses, number of restaurants, number of health facilities, number of amusement parks, and number of educational facilities are the independent variables with symbols.

Proposed Methodology

Understanding the issue is critical before moving on to the methodology. The issue is in the development of a hypothesis feature that can predict the target value based on the data provided in the training section. Then look at or evaluate the prediction on the data's testing section. The information provided here pertains to the house price and the features that go along with it. As a result, developing a computer that can learn data features and accurately predict price is a difficult challenge. With the aid of this model, the real estate builder's society will be able to easily predict the price of property, houses, and other items based on their characteristics.

This thesis' data set comes from Kaggle's Housing Data Set Knowledge Competition. This study aims to predict the house price (residential) in Bangalore, using a simple data collection. As a result, the data was obtained from the Kaggle Housing Datasets. The dataset's details are as follows: It has 81 explanatory variables, also known as features or characteristics. The target value is the last variable; in this case, it is called Sale Price, which is the actual price of the property. When the computer predicts the price, it will be compared to the actual value, and the mean error will be measured, giving the model's accuracy rate.

The data set could include information about the houses' various details. With the variables representing every aspect of residential homes in Bangalore, the researcher is faced with the task of predicting each home's final price. Now, using the panda’s library in the Python framework, import the data set and analyses it. Examine all of the house's features that are relevant to the dependent objective. Check for missing values and fill in all missing values by taking the median of all the values for that attribute when analyzing and visualizing the data..

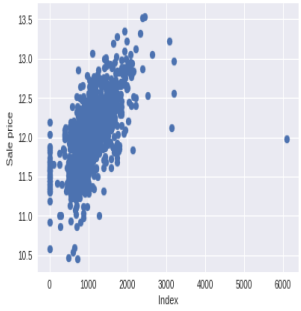
Pseudo-code for the Algorithm

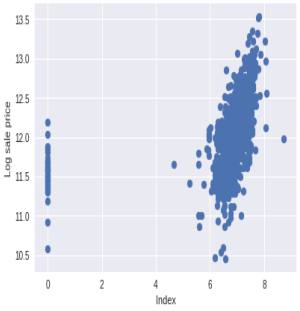
Step 1: Obtain the data set in the proper format.
Step 2: When the data set is being cleaned, look for any missing values.
Step 3: Delete the categorical data using the one hot encoder.
Step 4: Function selection using a heatmap or matrix correlation.
Step 5: Break the dataset into two sections, one for training and the other for testing.
Step 6: Match the regression methodology using the Training data and then validate the data with testing data.
Step 7: Compare the results of the accuracy test.

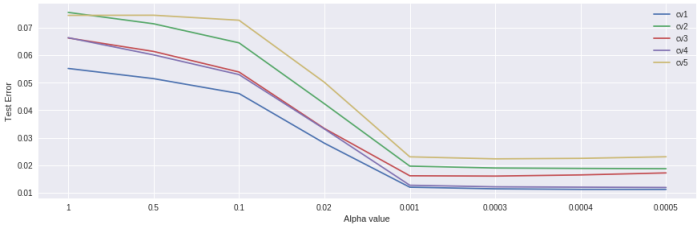
Results

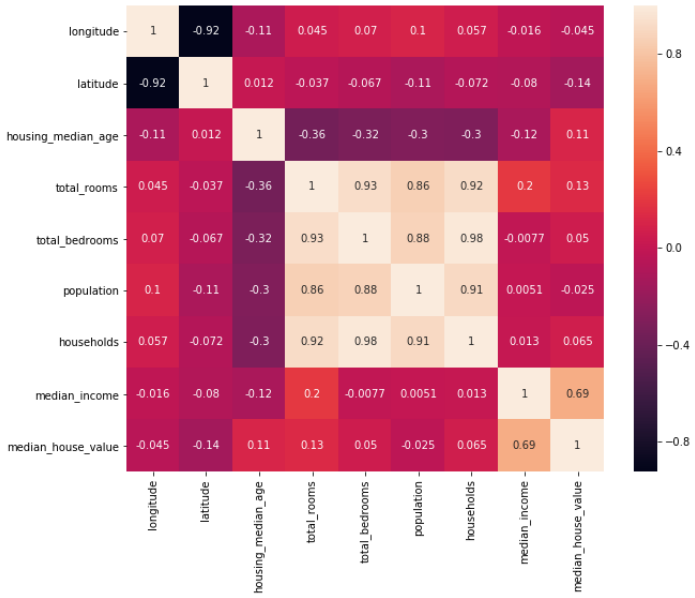
Below are the resulting mean error and residual outputs generated by the respective methods-

I. Multiple Linear Regression

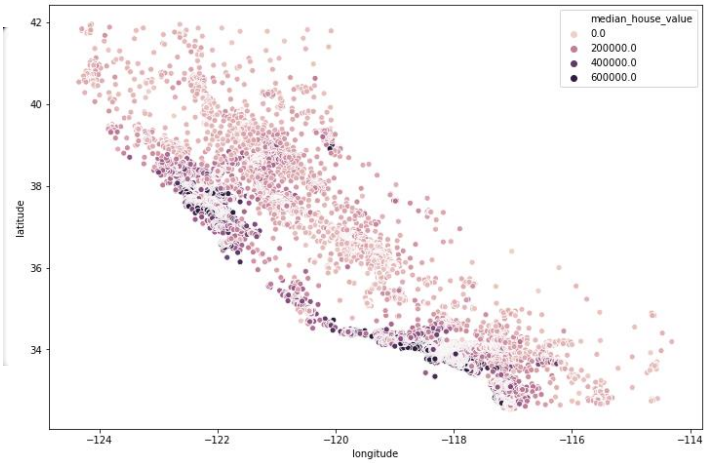








II. Keras Regression



Conclusion

It was able to estimate the predictive performance of a model using a variety of analytical and graphical methods.

Price models for housing come in a range of shapes and sizes. The models also helped categories which housing characteristics were most strongly linked to price and could explain the majority of the price difference. Furthermore, by accounting for the effect of various regression techniques, it was possible to improve the prediction accuracy of the models.

Easy and multiple linear regression, random forests, and gradient boosting for predictors were all used in this analysis. As a performance metric criterion, the models were evaluated and calculated using median absolute error and median percent error. The importance of each predictor in illuminating price variation for a specific collection of housing features was another major aim of this study. Overall, the findings provide useful information about the causes of different features on house prices, as well as their interpretation.

References

1. Wikipedia.com
2. “The Applied Hedonic Pricing Model of the Housing Market of Savannah and Its Savannah Historic Landmark District,” 22 in The Review of Regional Studies, vol. 39, by Rochard J. Cebula.
3. “House Price Determination: A Decision Tree Approach,” by Gang-Zhi Fan, Seow Eng Ong, and Hian Chye Koh.
2301–2315 in Urban Studies 43.12, 2006.
4. “Housing price dependent on ge-netic algorithm and help vector machine,” Expert Systems with Applications Gu Jirong, Zhu Mingcang, and Jiang Liuguangyan.
5. Hasan Selim, “House Price Determinants in Turkey: Hedonic Regression versus Artificial Neural
Network.” Expert Systems.