

**MACHINE LEARNING PROJECT REPORT**

**HOUSING PRICES PREDICTION**

**Submitted to: Submitted by:**

Mrs. Juhi Jain Group – 31

Sarthak Arora (2K18/CO/325)

Saurabh Singh (2K18/CO/329)

# INDEX

|  |  |  |
| --- | --- | --- |
| S.No. | Title | Page No. |
| 1. | Motivation | 3 |
| 2. | Objective | 3 |
| 3. | Literature Review | 3 |
| 4. | Experimental Design | 4 |
| 5. | Research Methodology | 6 |

# Motivation

* Growing unaffordability of housing – one of the major challenges for metropolitan cities around the world.
* Gain better understanding of the commercialized housing market that we are currently facing, we want to figure out what are the top influential factors in the house pricing.

# Objective

To provide people with better predictions about house pricing based on different factors on which the house prices depend so that they can plan their finance well.

# Literature review

# Recently, a few writers’ scopes for finding the best properties for the customers came along with various technologies.

* Li Li and kai-Hsuan Chu (2017) studied various algorithms such as Backpropagation neural network (BPN) and Radial basis functional (RBF) neural networks.
* Often the location's environmental conditions decide what kind of price we can expect for different types of houses, Manjula (2017) presents various important features to use when forecasting property prices with good precision using a regression model.
* Nihar Bhagat, Ankit Mohokar, Shreyash Mane (2016) studied linear regression algorithms for prediction of the houses. The goal of the paper is to predict the efficient price of real estate for customers with respect to their budgets and priorities.
* Hujia Yu, Jiafu Wu (2014) used classification and regression algorithms.
* Before this, Raghunandhan mentioned the basic data mining concepts of how it works and supporting algorithms for the purpose of prediction.

# Experimental Design

## DATASETS

Boston House Prices Dataset: The dataset used in this project comes from Kaggle. This data consists of 489 rows (data) and 4 columns (features).

The features can be summarized as follows:

* Average number of rooms among homes in the neighborhood (ANR).
* Percentage of homeowners in the neighborhood considered as ‘working poor’ (WPH).
* Ratio of students to teachers in primary and secondary schools in the neighborhood (RST).
* Median value of owner-occupied homes in $1000s (PRICE).

## REASEARCH VARIABLES

**Independent variables:** Independent variables are the ones which through we are trying to explain the value or effect of the output variable (dependent variable) by creating a relationship between an independent and dependent variable.

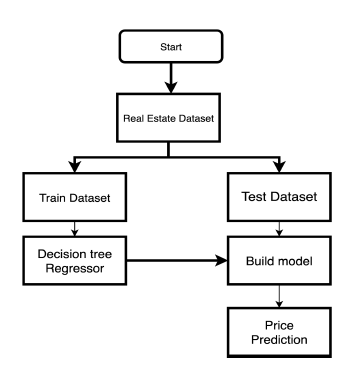
We work on the following independent variables in this project.

* Average number of rooms among homes in the neighborhood (ANR).
* Percentage of homeowners in the neighborhood considered as ‘working poor’ (WPH).
* Ratio of students to teachers in primary and secondary schools in the neighborhood (RST).

**Dependent variables:** Dependent variables are the variable in dataset which holds the phenomena which we are studying.

* Median value of owner-occupied homes in $1000s (PRICE).

## FLOW CHART



# Research Methodology

## DATA PREPROCESSING TECHNIQUES:

1. **Identifying and handling of any missing values:** We check for any missing values in the dataset. Missing values can cause malfunctioning of the model. On checking for missing values in dataset, we got none missing values.
2. **Detection and handling of outliers/anomalies:** Outliers are extreme values that deviate from other observations on data, they may indicate a variability in a measurement, experimental errors or a novelty. In other words, an outlier is an observation that diverges from overall pattern of the sample data.

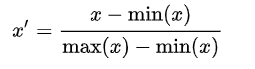
We applied Elliptical Envelope as an outlier technique in this model. It draws an imaginary ellipse around the dataset. Points outside this ellipse are outliers.

1. **Feature Scaling - Standardization and Normalization**

Feature scaling is a method wont to normalize the range of independent variables of data. In data processing, it is referred to as data normalization and is usually performed during the data preprocessing step. It is important to get information on one scale sometime and hence it can be achieved through it.

1. **Rescaling (min-max normalization)**

Rescaling is also referred as min-max normalization, is that the simplest method and consists in rescaling the range of features to scale the range in [0, 1] or [−1, 1]. Selecting the target range depends upon the nature of the data. The general formula for a min-max range of [0, 1] is given as:

{\displaystyle x'={\frac {x-{\text{min}}(x)}{{\text{max}}(x)-{\text{min}}(x)}}}

where {\displaystyle x} is an original value and {\displaystyle x'} is the normalized value.

1. **Standardization (Z-score Normalization)**

Standardization makes the values of every feature within the dataset have zero-mean (when subtracting the mean within the numerator) and unit-variance. The method of calculation is to determine the distribution mean and standard deviation for each feature. Next, we subtract the mean from each feature. Then we divide the values of each feature by their standard deviation.

{\displaystyle x'={\frac {x-{\bar {x}}}{\sigma }}}

Where {\displaystyle x} is the original feature vector,{\displaystyle {\bar {x}}={\text{average}}( {\displaystyle {\bar {x}}={\text{average}}(x)} is the mean of that feature vector, and {\displaystyle \sigma } is its standard deviation.

1. **Splitting the data set into training set and testing set.**

We split our dataset into training and testing dataset with 80% data as training set while the rest as the testing set. While training set data use to train and validate model, testing set data used to measure performance of the model.

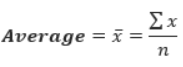
## DATA ANALYSIS TECHNIQUES:

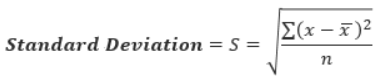
**Decision Tree Regression**

Decision tree regressor observes features and builds regression or classification models in the form of a tree structure. Breaks down dataset into smaller and smaller subsets while parallelly developing an associated decision tree. The Root node (topmost decision node) corresponds to the best predictor. It can handle both categorical and numerical data.

A decision tree builds from root in top-down order and involves partitioning the dataset into subsets of instances with similar values (homogenous). Homogeneity is decided on the basis of standard deviation of that numerical sample. Totally homogeneous numerical data corresponds to zero standard deviation.









* + Standard Deviation (S): for tree building (branching).
  + Coefficient of Deviation (CV): used to decide when to stop branching.
  + Count (n): can also be used for branching.
  + Average (Avg): the value in the leaf nodes.

**Grid Search optimization:**

Grid Search CV is a way to deal with parameter tuning that efficiently produces and evaluates a model for every mix of calculation parameters stored in a grid. Grid Search CV in this algorithm is used to evaluate the best-fit value for max-depth, using which the decision tree is formed.

## VALIDATION TECHNIQUES:

**K-Cross Validation Techniques**

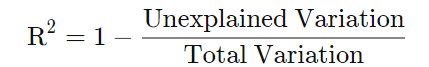
We used k-Cross validation technique to validate our model. It is a procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter namely k that refers to the number of groups in which given data sample is to be split into. As a results, we get a less biased or less optimistic estimate of the model skill.

The general procedure is as follows:

1. Shuffle the dataset randomly using ShuffleSplit() method.
2. Split the dataset into k groups
3. For each unique group:
   1. Take the group as test data set and remaining as training data set
   2. Fit the model on the training set and evaluate it on the test set
   3. Retain the evaluation score of model and then discard the model
4. Summarize the skill of the model using the sample of model evaluation scores

## PERFORMANCE MEASURE:

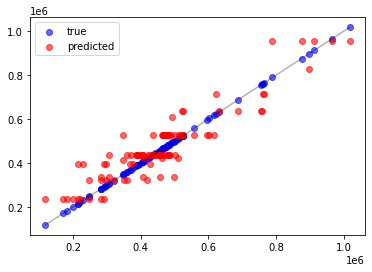
R - Squared Parameter(R2): It is a measure of the linear relationship between X and Y. It explains the proportion of the variance in the dependent variable that’s predictable from the independent variable.



More the R2 value better the performance. For this model, we got 83.73% R2 value which is pretty good.

RESULTS:

We get the performance of 83.73% on this model which is pretty good. On applying this model on test data, we get the following plot of predicted Vs actual prices:



VALIDITY THREATS:

This model can predict the price for any house, but with the inclusion of more features or preferences and exclusion of some features can disturb the quality of the model and can be less effective in that case. As it is up to the buyer’s choice and preference but we tried to include all more important and effective features in the pricing of a house.

CONCLUSION:

In this paper, the Decision tree machine learning algorithm is used to construct a prediction model to predict potential selling prices for any real estate property. The features included in the dataset, influence people’s decision while purchasing a property. Some of these features are not mostly included in the datasets of other prediction systems, which makes this system different. The system provides 83% accuracy while predicting the prices for the real estate prices.

FUTURE SCOPE:

In the future, a comparative study of the model’ predicted price and the housing prices from various property-based websites can be performed for the same user input. This can also help the user to compare the prices predicted by the model and the prices shown on the websites which can give the user a very good insight to the buyers regarding the prices of a particular property. The predicted prices can also be used to recommend real estate properties according to the buyers’ choice. The current dataset only includes city of Boston. Expanding it to other cities and states of India is the Future goal is to expand this model to different states and cities of India. The inclusion of factors like the presence of neighborhood amenities such as hospitals, schools within 1 km from the given location, in the procedure of making housing prices predictions increases the valuation of real estate property.

REFERENCES:

1. R. J. Shiller, “Understanding recent trends in house prices and home ownership,” National Bureau of Economic Research, Working Paper 13553, Oct. 2007.

2. Pow, Nissan, Emil Janulewicz, and Liu Dave Liu. "Applied Machine Learning Project 4 Prediction of real estate property prices in Montréal." Course project, COMP-598, Fall/2014, McGill University (2014).

3. Aaron Ng, Dr. Marc Deisenroth ..“Machine Learning for a London Housing Price Prediction Mobile Application” (2015)

4. R Manjula, Shubham Jain, Sharad Srivastava and Pranav Rajiv Kher..“Real estate value prediction using multivariate regression models” (2017)

5. Nihar Bhagat, Ankit Mohokar, Shreyash Mane "House Price Forecasting using Data Mining" International Journal of Computer Applications,2016.

6. Yu, H., and J. Wu. "Real estate price prediction with regression and classification CS 229 Autumn 2016 Project Final Report 1–5." (2016).