Shahjalal University of Science and Technology  
Department of Computer Science and Engineering



Submitted by :

Name : Bulbul Istehaque

Reg : 2014331012

Name : Tonmoy Bipro Das

Reg : 2014331041

Submitted to :

Ms Ayesha Tasnim

Assistant Professor

SUST

**House price prediction**

**Introduction**

In Machine Learning ‘House price prediction’ is a supervised regression problem .Because , space & price is not discrete .

**Data set**

We used “The Boston Housing Dataset” for this problem. This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. The dataset is small in size with only 506 cases. The data was originally published by Harrison, D. and Rubinfeld, D.L. `Hedonic prices and the demand for clean air', J. Environ. Economics & Management, vol.5, 81-102, 1978.

For each model, we split dataset into two part,

* Training data - 90%/ 85%/80%
* Testing data - 10%/ 15%/20%

**Used Model and Result**

1. **linear regression**
2. Linear regression is commonly used for predictive analysis. Simple Linear regressionis a statistical method that allows us to summarize and study relations between continuous variables Here, scalar response is price and other features are explanatory.

The accuracy of this model is 73.203% [test=.1]

72.8% [test=.15]

1. **Ensemble**

**Gradient boosting is a**[**machine learning**](https://en.wikipedia.org/wiki/Machine_learning)**technique for**[**regression**](https://en.wikipedia.org/wiki/Regression_(machine_learning))**and**[**classification**](https://en.wikipedia.org/wiki/Classification_(machine_learning))**problems, which produces a prediction model in the form of an**[**ensemble**](https://en.wikipedia.org/wiki/Ensemble_learning)**of weak prediction models, typically**[**decision trees**](https://en.wikipedia.org/wiki/Decision_tree_learning)**. It builds the model in a stage-wise fashion like other**[**boosting**](https://en.wikipedia.org/wiki/Boosting_(meta-algorithm))**methods do, and it generalizes them by allowing optimization of an arbitrary**[**differentiable**](https://en.wikipedia.org/wiki/Differentiable_function)[**loss function**](https://en.wikipedia.org/wiki/Loss_function)**.**

* + We used Gradient tree boosting and accuracy is 92.155% .
  + Accuracy increases as n-estimators increase .
  + For best accuracy maximum depth is 5 & minimum sample split is 2 .

1. **SVM**

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimentional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.The Support Vector Regression (SVR) uses the same principles as the SVM for classification, with only a few minor differences. In the case of regression, a margin of tolerance (epsilon) is set in approximation to the SVM which would have already requested from the problem.

The accuracy of this model is -7.173% .