Report on CPU Price Prediction

Abdullah Al Sefat (14.02.04.009)
Rashik Islam (14.02.04.012)

1 Problem Description

There are many variants of CPU available in the market. The prices of these CPUs range from 50 pounds to upto more than 2000 pounds. Customers want to buy CPUs according to their usage. The variance of price is very high. It will be helpful if we can get an estimated price according to needs.

2 Datasets Dedscription:

The dataset that we have prepared consists of 11 columns in total. 10 of them are feature variables and one of them is the target. The features are different attributes of a processor and the target is price in great britain pounds.

2.1 Size of the Dataset

- 104 instances
- 10 attributes
- 1 target

2.2 Data Definition

Table 1: Data Definition

variable	Definition
Brand	The manufacturer of the processor
Cspeed	Clock speed of the CPU in Giga Hertz
Cache	The size of L3 Cache in Mega Bytes
HT	CPU has Hyperthreading or not (For each physical thread a logical thread is mapped)
TB	CPU can TurboBoost or not (Increase Clock speed when under load)
Unlocked	CPU is overclockable or not
Lithography	Size of the transistors in nano metres
SSC	Single core Cinebench R15 performance score
MSC	Multi core Cinebench R15 performance score
Core	The number of Physical Cores
Price	The retail price of the CPU in Great Britain Pounds

3 The Machine Learning Models Used:

- Linear Regression
  - Linear regression model tries to fit the data points on a line in an n+1 dimensional space where n is the number of features.
- Support Vector Machine
  - Support vector machine uses a set of points called support vector point. There is a decision function, which is evaluated by these vector points.
- Decision Tree
  - A tree is constructed. Leaf contains the result. Starting at the root we traverse down by taking decisions at each node. at each node there is a condition corresponding to a child.
- Random forest
  - Random forest regression is an ensemble machine learning algorithm. Random forest consists of large collection of decorrelated decision trees. Random forest takes the decisions of those decision trees and gives the mean as result. This increases the accuracy of the result.
- Gradient Boostig
  - Gradient boosting is also an ensemble machine learning algorithm. Ensemble algorithms use multiple models and then combines their reults by takig average using weighted voting. In Gradient Boosting at first there is an initial model. At first we learn the model. Then calculate the error residual. The we fit a model fitting the error residual. Then we combine this model the previous model. Repeat the same procedure. Now we can see the accuracy has improved.

4 Performance scores of the models

Now let us see how the model performs. We have taken r^2 score, Mean Absolute Error, Mean Squared Error and Root Mean Squared Error in consideration.

Table 2: Performance metrices of various models

Model	R2	MAE	MSE	RMSE
Linear Regression	0.749710031586	111.918531014	32710.9397139	169.784431906
Support Vector Machine	-0.143970219676	238.749486398	236327.703043	456.302788212
Decision Tree Regression	0.78684953541	91.1633698413	39837.0304257	178.282761247
Random Forest Regression	0.83297716177	86.3826349206	35410.4496548	169.583314544
Gradient Boosting Regression	0.835894544577	81.3627427446	35351.0551429	162.402625805

5 Discussion

From the performance metrices we can see that ensemble models work the best. Gradient Boost Regression has the best scores. Other models have decent score. But SVM has the worst results. It can be that it is due to the way the variables were scaled.