Laptop Price Prediction

Group : A1
Lab Group Number : 04

Submitted By:
Syed Sajid Mahboob 14.02.04.10
Tanjila Hossain 14.02.04.020

1 Problem

This project is a Regression problem that predicts the price of laptop based on some important features.

2 Problem Description

Laptop prices vary from laptop to laptop depending on laptop components.
In the project the data-set that has been used includes the components like type of the processor, memory, hard drive, optical drive, speed etc and their corresponding prices of different laptops.
After analyzing the data-set this project is trying to predict the prices for the new feature values.

3 Data-set Description

The data-set consists of , 16 columns among which 15 are features and the last column is the target.

1. **dual-core**: Laptop processor's type.
2. **quad-core**: Laptop processor's type.
3. **core-i3**: Laptop processor's type.
4. **core-i5**: Laptop processor's type.
5. **core-i7**: Laptop processor's type.
6. **memory-GB**: RAM size.
7. **storage-GB**: Memory storage of the laptop(hard drive)in GB.
8. **display-inch**: Laptop screens/ Display.

9. **graphics-card**: Laptop's graphic card. Assuming for Intel(1), AMD(2), Nvidia(3).
10. **battery-cell**: Number of cells of the battery.
11. **audio-speakers**: Audio speaker of the laptop.
12. **optical-drive**: Presence of laptops optical drive (DVD or CD drive)-if exists then 1 otherwise 0.
13. **cache-MB**: Cache size(the size of the data store).
14. **speed-max-GHz**: Speed of the laptop's processor(GHz).
15. **processor-gen**: Laptop processor generation(processor number).
16. **price**: Price of the laptop.

4 Used Models

5 models are used in this problem to see which one gives the better R^2$score.Scoreforeachmodelistestedi$
$FoldCrossValidationwithk = 5wedividedourdatasetinto5slicesandranregressionondifferentslicese$

4.1 Linear Regression

It is one of the most widely known modeling technique. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.

Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line).

It is represented by an equation Y=a+b*X + e, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

4.2 Decision Tree Regression

Decision tree builds classification or regression models in the form of a tree structure.It doesnt require any data transformation. It means that we dont have to spend more time preprocessing the data. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

A decision tree splits the input features in several regions and assigns a prediction value to each region. The selection of the regions and the predicted value within a region are chosen in order to produce the prediction which best fits the data. Where for best fit we mean that it minimizes the distance of the observations from the prediction.

4.3 Random forest Regression

The Random Forest is one of the most effective machine learning models for predictive analysis. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting mean prediction (regression) of the individual trees.

The random forest model is a type of additive model that makes predictions by combining decisions from a sequence of base models. More formally we can write this class of models as:

- g(x)=f0(x)+f1(x)+f2(x)+...

where the final model g is the sum of simple base models fi. Here, each base classifier is a simple decision tree. This broad technique of using multiple models to obtain better predictive performance is called model ensembling. In random forests, all the base models are constructed independently using a different subsample of the data.

4.4 AdaBoost Regression

The core principle of AdaBoost is to fit a sequence of weak learners (i.e., models that are only slightly better than random guessing, such as small decision trees) on repeatedly modified versions of the data. The predictions from all of them are then combined through a weighted majority vote (or sum) to produce the final prediction.
An AdaBoost [1] regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction. As such, subsequent regressors focus more on difficult cases.

This class implements the algorithm known as AdaBoost.R2

4.5 SGD (Stochastic Gradient Descent) Regression

Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to discriminative learning of linear classifiers under convex loss functions such as (linear) Support Vector Machines and Logistic Regression.
The class SGDRegressor implements a plain stochastic gradient descent learning routine which supports different loss functions and penalties to fit linear regression models.

1. The advantages of Stochastic Gradient Descent are:
   - Efficiency.
   - Ease of implementation (lots of opportunities for code tuning).
2. The disadvantages of Stochastic Gradient Descent include:
   - SGD requires a number of hyper parameters such as the regularization parameter and the number of iterations.
   - SGD is sensitive to feature scaling.

Model	**Score**
Linear	0.92109
Random Forest	0.87083
AdaBoost Regressor	0.87625
Decision Tree Regressor	0.66195
SGDRegressor	0.25182

5 Conclusion

As we can see 4 out of 5 models give R$^{2} scorewhicharealmostclosetoeachother.ButincaseofSGDRegres$
$10.000) butourdata - setcontainsonly <= 150 data.$