Laptop Price Prediction by Regression Models

Annal Ali Albeeshi, Aesha Bakheet Aljohani

Supervisor the project:

Mohamed Eldeeb, kinza waqar

Abstract:

The (COVID19) pandemic came to consolidate the concept of the importance of technology, including computers, laptops, tablets and mobile phones of all kinds; One of the most important changes that resulted from the pandemic for nearly two years is the change in the method of education, which has become and is still online. That is why we wanted to shed light on computers through their prices, specifications and brands through Regression models.

Design:

Regression models were used to predict the prices of mobile devices, by integrating of two dataset**.**

Data Description:

The csv file contains of 1853 rows and 11 columns. The description of each field is as below:

No	Column name	Data type	Description
1	Company	String	Laptop Manufacturer
2	Product	String	Brand and Model
3	TypeName	String	Type (Notebook, Ultrabook, Gaming, etc.)
4	Display	Numeric	Screen Size
5	CPU	String	Central Processing Unit (CPU)
6	RAM	Numeric	Laptop RAM
7	Memory	String	Hard Disk / SSD Memory
8	Operating System	String	Operating System
9	Weight	String	Laptop Weight
10	Price-USD	Numeric	Price
11	Warranty	String	

Methodology:

1. **Pre Processing.**
2. **EDA Analysis.**
3. **Algorithms.**

1- Pre Processing:

i. Data Collation.

Tow datasets have been selected from kaggle, the first dataset is and it consists of 1303 rows and 13 columns (1). The second dataset consists of 550 rows and 10 columns (2).

ii. Data cleaning

Through this step, it was made sure that there are no duplicate rows in both tow datasets, and the absence of missing values, also the data in some columns was modified, for example: in the two columns OpSys and Operating System, which contained Windows 7, Chrome OS , and 64 bit Windows 10 Operating System ...etc, became win10, chrome, and 64-bit win10. Create new columns, convert the data type for some columns, convert the currency and make it one currency, which is USD, change the names of some columns until the datasets are combined.

iii. Data target

By creating a new data frame called final_laptop.csv, It consists of 1853 rows and 20 columns.

iv. Data cleaning for data target

At this stage we did some processing on the new set of data that we obtained after the merger process, including: droping colunams, check for missing data (nan value), where was found in Warranty 1303 missing data, TypeName 550 missing data, Weight 550 missing data, RAM 1 missing data. And check for duplicates or unnecessary data, where was found our dataset contains 36 duplicate rows , and by removing these rows we have 1817 rows, where they were previously 1853 rows., check for outliers, where was found our datasets do not contain outliers, through the box plot .

2- EDA Analysis:

3- Algorithms.

Regression models

1- Random Forest Regression model.
2- Linear Regression model.

Models building steps :

1- Label Encoder:
2- Train model
3- Model evalulation
4- Experiments
5- Comparison of the performance of the models

1- Label Encoder:

Through this step a label encoder can be done for all columns, including the categorical.

2- Train model

The model is trained by dividing the dataset into two sets: training set 80% and testing set 20%. Through the training set, the model is trained, and through the test set, the prediction is calculated by *score*. As shown in Table 1 the results for each model in each training set and testing.

Table 1: prediction results

SCORE	Linear Regression model	Random Forest Regression model
Training set 80%	0.54	0.97
Testing set 20%	0.50	0.96

3- Model evalulation on testing data:

Table 2: Model evaluation results

	Linear Regression	Random Forest Regression
Mean Squared Error (MSE)	283926	83
Mean Absolute Error (MAE) =	361	19389
Root Mean Square Error (RMSE)	532	139
R^2 Score	0.50	0.96

4- Experiments:

This step is one of the optimization methods the model, it will be experiments for the Linear Regression model because of the low score results, and as shown in the table 3 experiments and the results.

Table 3: Experiments and the results

score	select some of the columns	Polynomial Feature	adding interaction terms
Train R^2	0.44	0.56	0.57
Validation R^2	0.38	0.53	0.53

Conclusion:

We note that the prediction results for the Linear Regression model are low even after conducting experiments, as there is a slight increase in the results, but in general the results are not satisfactory or high, like the Random Forest Regression model where the prediction results were in training = 97 and testing = 96.

Future work

1- Scrap data form website.
2- Explore different models.

References

1- https://www.kaggle.com/muhammetvarl/laptop-price
2- https://www.kaggle.com/asifraza14/laptop-price-prediction-using-specifications