# **Introduction (**1. A clear introduction, motivation, a description of the problem domain, and an explanation of how the project's goals are justified using Prediction / Classification algorithms.)

Laptops have become an indispensable part of everyone’s lives. They are beneficial for quickening the pace when it comes to any task. They serve the purpose of communication through online interactions and streamline other operations (TechGuideSpot, 2020).

Are various advantages of using a laptop computer as there are portable which can carry very easily from one place to another, lightweight being light in weight, you can use the laptop anywhere, small in size, offline operations, no cables (Anon, n.d.)

To forecast the price of a laptop, this model is aimed at determining a laptop’s price by considering factors like its processor, RAM, memory type, brand, weight, CPU, and type of notebook among other things. With motivation to analyse characteristics that make the laptop cheaper or more expensive, or even have a median value according to characteristics. And see how this can be analysed using machine learning models. which can be a good tool for customers who want to buy, who can get a sense of the value of the product they want and can also help sellers set competitive prices based on market trends.

The problem is to analyze the price taking into account all the characteristics and attributes that involve the pricing of a product, such as selecting which characteristics have a strong influence when it comes to placing value on the object. The task is to develop a model that can effectively comprehend this relationship and provide an accurate representation of it.

# Data Understanding

For this project the dataset that was chosen is named “Laptop Price” available on Kaggle. Using the “df.shape” and “df.info” functions, is it possible to see that the data size consists of 1303 and 13 columns, of which eleven columns are represented by ‘object’ data type, and only

three column with numerical data, being consisted of float64 and int64 type.

By using the “df.describe” to obtaining the summary statistics for this dataset. We can the summary statistics for only numerical data, which are laptop\_ID, Inches and Price\_euros.

A screenshot of a computer

Description automatically generated

Figure 1. Summary (describe).

The measures of dispersion evaluate how scattered the collected data are. They are standard deviation, variation and interquartile range. If the value of the standard deviation is high, it means that the set of data is very spread out. In this case, the result of standard deviation is lower than the mean. So, it’s mean the set not very spread out.

# **Data preparation**

With the function of “df.head” it was possible to observe that in the categorical variables it was necessary to make some modifications.

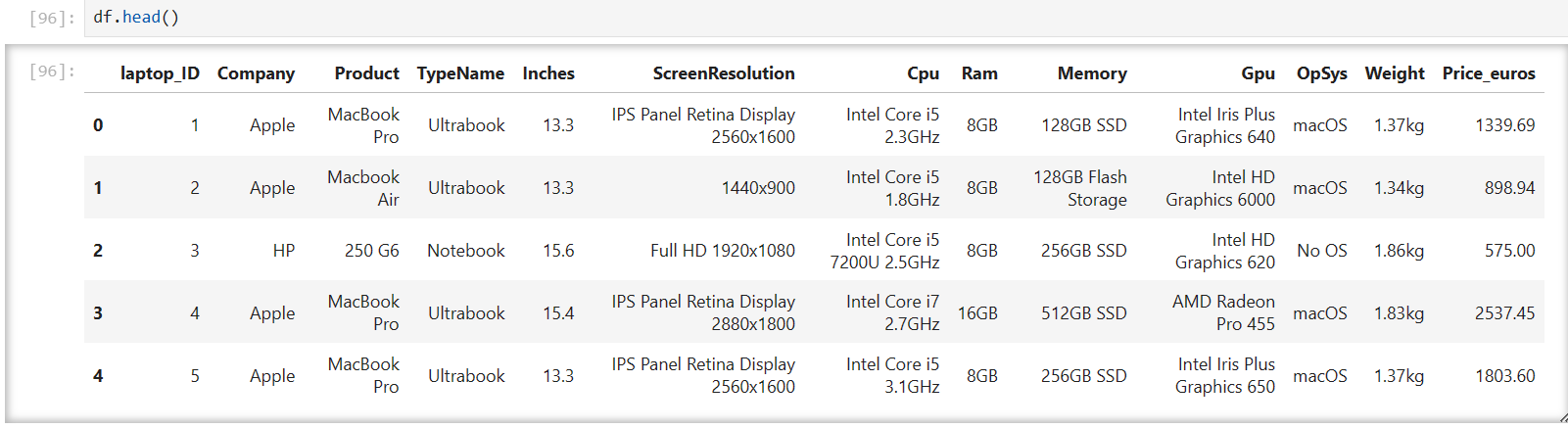


Figure 2. df.head

Starting by removing str from the RAM, Weight, Cpu\_ frequency columns, such as, GB, kg, GHz. as well as changing make replacements of dtype of the columns. Ram was transformed into int32, weight into float32. And the same thing happened on screenR, Cpu\_frenquency.

To extract information or unique values from some columns that have variables that need to be accounted for individually, we used the lambda function to a single column and dropout the first columns. Like as you can see in ScreenR and Cpu. New columns is defined to ‘ScreenR’, appearing two new columns ‘screen\_width’ and 'screen\_height'.

A screenshot of a computer code

Description automatically generated

Figure 3 Example of lambda application.

Had to put some memory values in the same value scale, because some are in GB and others in TB. So, it's important to put the values on the same scale.

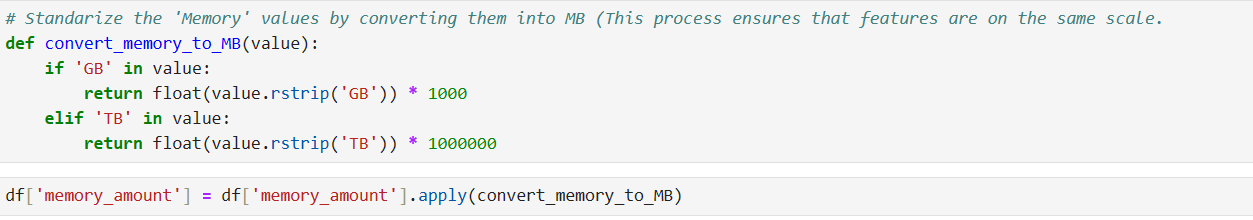


Figure 4: Standardize memory.

The ML algorithms deal with numerical data, so for this data we had to convert categorical feature to numerical data. For all categorical variables, we must transform them into columns that we can apply to the model. This is done by creating dummies, convert categorical variable into dummy/indicator variables. Each variable is converted in as many 1 to indicate every entry in a column and 0 if otherwise. Columns in the output are each named after a value. One-hot encode 'Company', 'TypeName', ‘memory\_type’, 'cpu\_brand', 'gpu\_brand' and 'OpSys' columns and delete the original columns. The data shape after that was 47 columns changed to a useful format.

# FEATURE SELECTION

to select columns that interfere with my target variable. A correlation function was used to observe which columns have the best correlation coefficients.

A screenshot of a computer screen

Description automatically generated

Figure 5. Correlation coefficients

By using the Filter absolute correlation values by threshold, we can select the features which has corr > 0.1, That way, we select only the feature that best correlates with the target feature. thus creating a new dataframe (new\_df) with 23 columns from filter selection.

# Modelling

First step:

Splitting the dataset into independent (features) and dependent (target) variables is a common first step. define what the target variable is, i.e. dependent, and which variables are independent. Here our x and y variables have been defined.

A screenshot of a computer code

Description automatically generated

Figure 6: Splitting the dataset

The data was split into train and test data. the algorithm model learns with the train while the test is used to check the model. After that, the data must be standardized to compress the data within a particular range of values.

Second step:

Selected the machine model to apply to the data. Having observed that the dependent variable is a variable containing decimal values. Regression algorithms are used to predict a continuous outcome (y) using independent variables (x). We constructed a Support vector regressor, Linear Regression and Random Forest as machine learning models.

* Support vector regression

Fitting our model to our training sets. For support vector kernel linear because that works better with the data, the R2 gave better values by using kernel linear. The linear model captures the underlying patterns in data more effectively.

A screenshot of a computer program

Description automatically generated

Figure 7: Svr model

By using different proportions of the data training. Was observed that test(0.25) gave the best values of R2 score 0.6982, but in general, the model only explains around 0.69-0.70 of the variance of the target variable, it is decent but the mean squared predicted values compared to the actual is quite large.

|  |  |  |
| --- | --- | --- |
| test(0.20) | R2 score: 0.6935206462758179 | Mean Squared Error: 146507.28156957857 |
| test(0.25) | R2 score: 0.698205132988984 | Mean Squared Error: 160505.27175493186 |
| test(0.30) | R2 score: 0.6959153395470743 | Mean Squared Error: 162831.21546469326 |

* Linear Regression

A screenshot of a computer program

Description automatically generated

Figure 8: LR

# REFERENCES

TechGuideSpot. (2020). *The Importance Of Laptops In Our Lives*. [online] Available at: https://www.techguidespot.com/the-importance-of-laptops-in-our-lives/?utm\_content=cmp-true [Accessed 21 Apr. 2024].

www.kaggle.com. (n.d.). *Laptop Price*. [online] Available at: <https://www.kaggle.com/datasets/muhammetvarl/laptop-price/data>.

Anon, (n.d.). *The Advantages and Disadvantages of Laptop Computer*. [online] Available at: https://quicklearncomputer.com/advantages-and-disadvantages-of-laptop/.