# Laptop Price Prediction

**Pattern Recognition and Machine Learning Project**

Pratik Patil *(202118023)*
*MSc DS, DAIICT*
Gandhinagar, Gujarat
prattikk.pattill@gmail.com

Tanya Jagyasi *(202118039)*
*MSc DS, DAIICT*
Gandhinagar, Gujarat
tanyajagyasi@gmail.com

April 29, 2022

**Abstract**

In today's modern world, Laptops are a very crucial part in the daily lives of every single person, whether it be a student, a businessman, a working professional or even even in fields apart from academics e.g. Gaming. As laptops are really compact and easy to carry anywhere it makes them a preferred choice over any other gadget. In this project we are going to apply various Machine Learning Algorithms to accurately predict the price of laptops based on several features.

## 1 Introduction

A laptop computer, sometimes called a notebook computer by manufacturers, is a battery- or AC-powered personal computer generally smaller than a briefcase that can easily be transported and conveniently used in temporary spaces such as on airplanes, in libraries, temporary offices, and at meetings.

Laptops are portable computer that encompasses a microprocessor, rechargeable battery, fold-down screen, keyboard and mouse.

There are numerous companies in the market that produce a variety of laptops from budget friendly and low quality to premium high quality laptops for different purposes like academic use, office use or gaming.

Each laptop has it's own different features and their price varies from brand to brand. There could be two laptops having the same specifications manufactured by two different companies but having different price. This happens because the pricing is completely dependent on the company, market conditions and the availability of resources.

But there is still need to find out the exact price of any given laptop irrespective of their brand, and this could be done if we have all the relevant information to train a machine learning model and predict the price of a laptop with any set of features.

## 2 Data Description

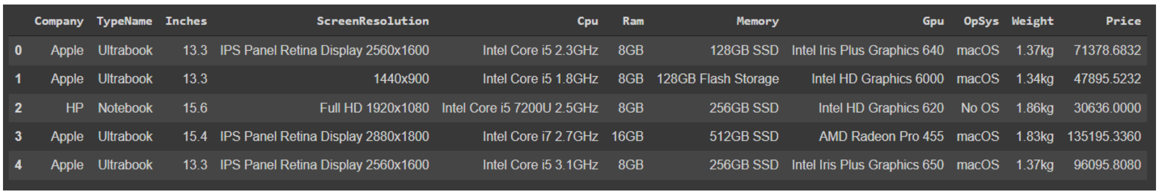
This section contains all the aspects of data from data description and collection.

This dataset is taken from Kaggle.com.

The attributes in the dataset are :

| Field Name | Description |
|---|---|
| Company | Name of Manufacturer |
| TypeName | Type of Laptop |
| Inches | Screen Size in Inches |
| ScreenResolution | type of Display and its resolution |
| Cpu | CPU name and type |
| Ram | RAM size in GB |
| Memory | Hard-Disk size and type |
| Gpu | GPU name and type |
| OpSys | Operating System |
| Weight | Weight of the laptop in Kg |
| Price | Price in Rupees(₹) |

The dataset consists of around 1300 rows and 12 columns.



| | Company | TypeName | Inches | ScreenResolution | Cpu | Ram | Memory | Gpu | OpSys | Weight | Price |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Apple | Ultrabook | 13.3 | IPS Panel Retina Display 2560x1600 | Intel Core i5 2.3GHz | 8GB | 128GB SSD | Intel Iris Plus Graphics 640 | macOS | 1.37kg | 71378.6832 |
| 1 | Apple | Ultrabook | 13.3 | 1440x900 | Intel Core i5 1.8GHz | 8GB | 128GB Flash Storage | Intel HD Graphics 6000 | macOS | 1.34kg | 47895.5232 |
| 2 | HP | Notebook | 15.6 | Full HD 1920x1080 | Intel Core i5 7200U 2.5GHz | 8GB | 256GB SSD | Intel HD Graphics 620 | No OS | 1.86kg | 30636.0000 |
| 3 | Apple | Ultrabook | 15.4 | IPS Panel Retina Display 2880x1800 | Intel Core i7 2.7GHz | 16GB | 512GB SSD | AMD Radeon Pro 455 | macOS | 1.83kg | 135195.3360 |
| 4 | Apple | Ultrabook | 13.3 | IPS Panel Retina Display 2560x1600 | Intel Core i5 3.1GHz | 8GB | 256GB SSD | Intel Iris Plus Graphics 650 | macOS | 1.37kg | 96095.8080 |

Figure 1: Dataset

This is a raw dataset with minimum specifications of the given specific laptops. After studying some other works on this topic we found that there are a few more features required for predictions, like, Pixels Per Inch (ppi) which could easily be calculated from the existing data that we had. And there was some other data pre-processing needed to be done before actually applying the regression algorithms on the data.

In the above mentioned features, there are some columns having String values which need to be converted to integers for better analysis. Also, there are some columns having combined string and integer values which need to be separated and the integer values are needed to be analyzed separately.

# 3 Data Pre-Processing

This dataset that

## 3.1 Checking for Duplicates and Null values

In data pre-processing, we have check for the dataset for whether it had any null values using the ‘isnull()’ function from the pandas library. Further we checked for duplicate values using the ‘duplicate()’ function. Where we found that there weren’t any duplicate or null values in the dataset and we can easily perform the further required operations for data pre-processing and exploratory data analysis in the further part of the project.

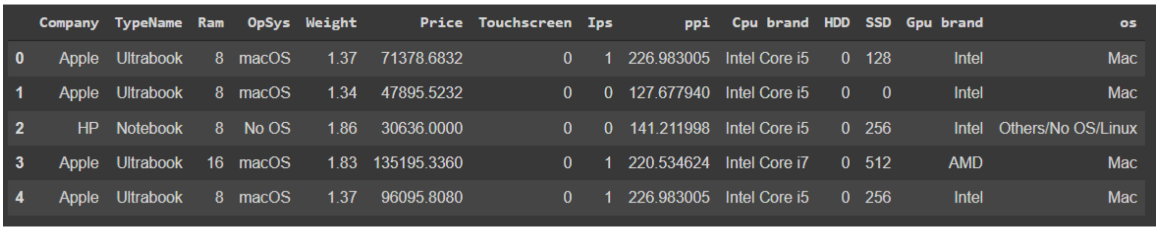
## 3.2 Changing Data Type

Further, we removed the ‘Unnamed’ serial number column from the dataset which was unnecessary. And then used string replacement function ‘str.replace()’ to replace the string values ‘GB’ and ‘kg’ to blank spaces from the ‘Ram’ and ‘Weight’ columns respectively to make it easier for converting these string columns to integer and float data type using the ‘.astype()’ function from the pandas library.

## 3.3 Separating data into separate columns

There were a few columns where string and numeric values were written together and needed to be separated in order to get better results, Columns like ‘ScreenResolution’, ‘TouchScreen’ and ‘IPS Display or not’.

Finally after pre-processing we got the following table to work upon:



| | Company | TypeName | Ram | OpSys | Weight | Price | Touchscreen | Ips | ppi | Cpu brand | HDD | SSD | Gpu brand | os |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Apple | Ultrabook | 8 | macOS | 1.37 | 71378.6832 | 0 | 1 | 226.983005 | Intel Core i5 | 0 | 128 | Intel | Mac |
| 1 | Apple | Ultrabook | 8 | macOS | 1.34 | 47895.5232 | 0 | 0 | 127.677940 | Intel Core i5 | 0 | 0 | Intel | Mac |
| 2 | HP | Notebook | 8 | No OS | 1.86 | 30636.0000 | 0 | 0 | 141.211998 | Intel Core i5 | 0 | 256 | Intel | Others/No OS/Linux |
| 3 | Apple | Ultrabook | 16 | macOS | 1.83 | 135195.3360 | 0 | 1 | 220.534624 | Intel Core i7 | 0 | 512 | AMD | Mac |
| 4 | Apple | Ultrabook | 8 | macOS | 1.37 | 96095.8080 | 0 | 1 | 226.983005 | Intel Core i5 | 0 | 256 | Intel | Mac |

Figure 2: Data after Pre-Processing

# 4 Exploratory Data Analysis

We have done the analysis of the data :

## 4.1 Bar Graph

The first bar graph shows that the distribution of laptops belonging to various companies.

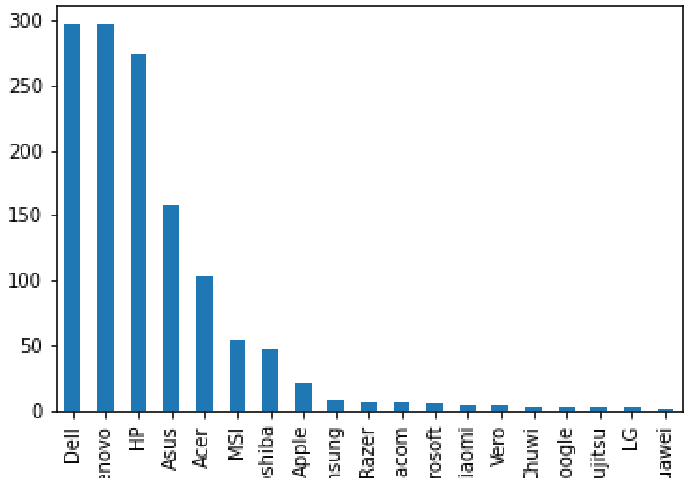


Figure 3: Distribution of laptop companies in the dataset

The second bar graph depicts the distribution of Laptop Type in our dataset.

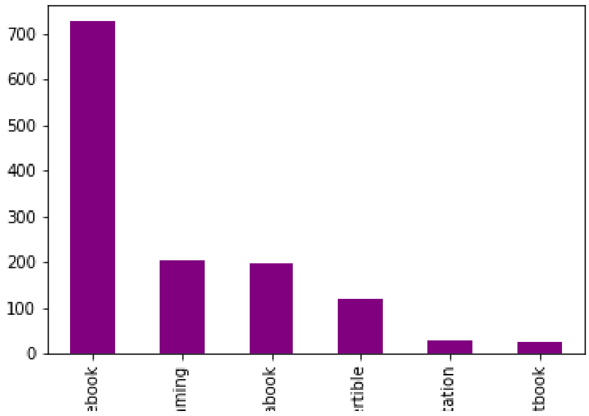


Figure 4: Distribution of Laptop Type

The third bar graph depicts the distribution of Laptop CPU Brands in our dataset.

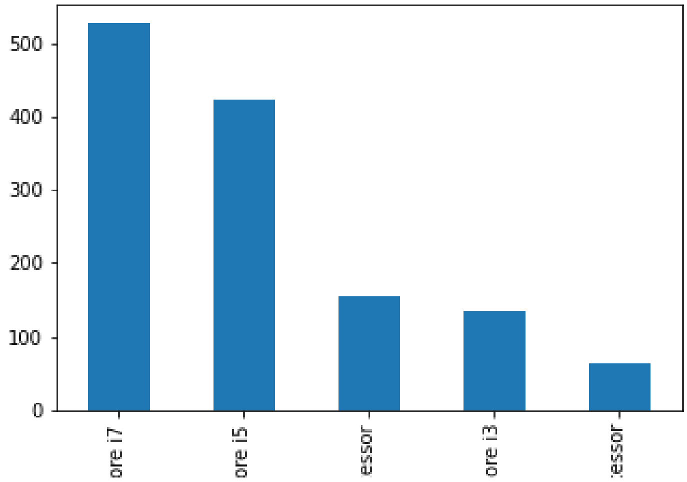


Figure 5: Distribution of Laptop Type

The forth bar graph depicts the distribution of Laptop Screen Size in our dataset.

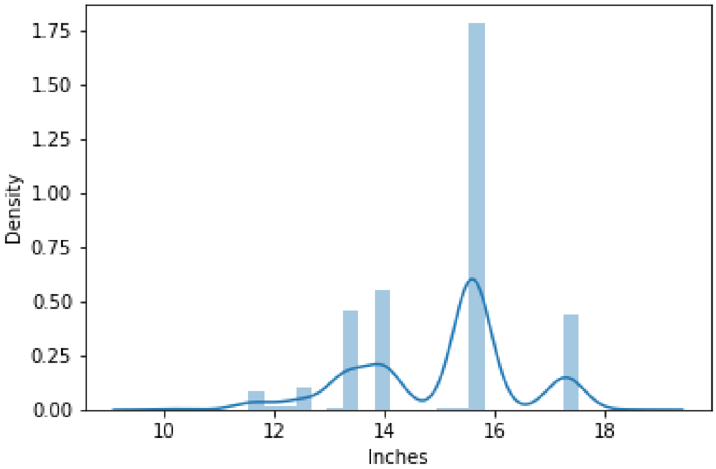


Figure 6: Distribution of Laptop Type

### 4.2 Correlation between all major features

Using sns.heatmap(), we have plotted a heatmap/corrplot which finds the correlation between each major feature from our data. The more negative the correlation become the deeper the colour of plot gets and lighter the colour, positive the value is.

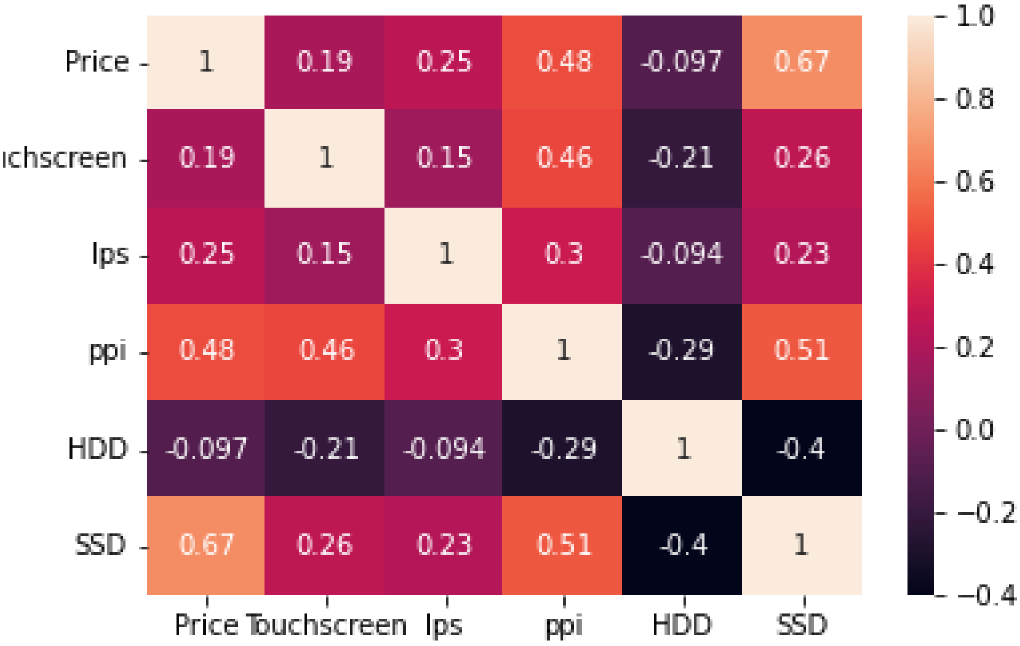


Figure 7: Heatmap/Corrplot

## 5 Machine Learning Models

We have used five Machine Learning regression algorithms for predicting the laptop prices.

The Regression models that we have used are:

1. Linear Regression
2. Ridge Regression
3. Lasso Regression
4. Decision Tree Regressor
5. Random Forest Regressor

## 5.1 Linear Regression

Linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line. It is commonly used for predictive analysis and modeling.

After applying Linear regression on our data we got:

***R-squared = 0.80732***

***Mean Squared Error = 0.07370***

***Root Mean Squared Error = 0.27149***

which look good as the mean squared error is very less and we can say that this model will give good prediction results if applied in real life scenario.

## 5.2 Lasso Regression

The lasso regression allows you to shrink or regularize these coefficients to avoid overfitting and make them work better on different datasets. This type of regression is used when the dataset shows high multicollinearity or when you want to automate variable elimination and feature selection.

After applying Lasso regression on our data we got:

***R-squared = 0.80718***

***Mean Squared Error = 0.07376***

***Root Mean Squared Error = 0.27159***

which look good as the mean squared error is very less and we can say that this model will give good prediction results if applied in real life scenario.

## 5.3 Ridge Regression

Ridge regression is a way to create a parsimonious model when the number of predictor variables in a set exceeds the number of observations, or when a data set has multicollinearity (correlations between predictor variables).

After applying Ridge regression on our data we got:

***R-squared = 0.81273***

***Mean Squared Error = 0.07163***

***Root Mean Squared Error = 0.26765***

which look better than Linear Regression as the mean squared error is very less and we can say that this model will give better prediction results if applied in real life scenario.

### 5.4 Decision Tree

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

After applying Decision Tree regressor on our data we got:

***R-squared = 0.84828***

***Mean Squared Error = 0.05803***

***Root Mean Squared Error = 0.24090***

which look better than Ridge Regression as the mean squared error is very less and we can say that this model will give better prediction results if applied in real life scenario.

### 5.5 Random Forest

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

After applying Random Forest regressor on our data we got:

***R-squared = 0.88734***

***Mean Squared Error = 0.04309***

***Root Mean Squared Error = 0.20760***

which look better than Decision Tree as the mean squared error is very less and we can say that this model will give better prediction results if applied in real life scenario.

## 6 Conclusion

After developing all the five models to predict laptop prices for any given specification, we found out that Random Forest regression performs the best for predictions as it has the highest R-squared score and a very low Root Mean Squared Error compared to the other four models that we applied.

For problems like this generally Random Forest gives the best results as it randomly selects data samples and uses multiple Decision Trees to train the model in a randomized way. Random Forest regeressor takes the average of all it's decision tree's to give the final prediction which is a good way to reduce variance in the predicted results. All this gives better predictions and makes Random Forest the best choice for regression model in this problem.

## References

[1] https://www.academia.edu/69591584/Laptop_Price_Prediction_using_Machine_Learning

[2] https://medium.com/analytics-vidhya/laptop-price-prediction-by-machine-learning-7e1211bb96d1

[3] https://nycdatascience.edu/blog/student-works/laptop-recommendation-system/

[4] https://dataaspirant.com/lasso-regression/

[5] https://www.statisticshowto.com/ridge-regression/