**INTRODUCTION**

The price of a laptop is influenced by multiple factors, including brand, specifications, and hardware configurations. With the growing variety of laptops in the market, accurately predicting their prices based on specifications can be valuable for consumers, manufacturers, and retailers. Machine learning (ML) models provide an efficient way to analyze complex relationships between laptop features and their prices.

In this project, we develop a machine learning-based regression model to predict laptop prices using a dataset containing 1303 data points. The dataset includes various attributes such as brand (Company), Type Name, Screen Size (Inches), Screen Resolution, Processor (CPU), RAM, Storage (Memory), Graphics Card (GPU), Operating System (OpSys), Weight, and Price. Several regression models were trained and evaluated, with the best-performing model selected based on R-squared and Mean Absolute Error (MAE) metrics.

**OBJECTIVE**

The primary objective of this project is to build an accurate laptop price prediction model using machine learning regression techniques. The specific goals include:

* Analyzing the impact of different laptop features on price based on their respective correlations and Exploratory Data Analysis (EDA).
* Fitting of several ML Regression models.
* Comparing the performance of Regression models.
* Selecting the best model based on R-squared and MAE values.
* Developing a web application which will take the laptop specifications as an input from the user and provide an estimate of that specified laptop as an output. The selected Regression model will work on the back-end to provide that prediction.

**METHODOLOGY**

* **Loading the dataset**.
* **Data Preprocessing** which includes:
  1. Handling missing values (if any)
  2. Encoding the categorical variables before feeding the data into the models.
* **Exploratory Data Analysis (EDA)** which includes:

1. Studying feature correlations to understand relationships
2. Visualize trends and distributions

* **Feature Selection & Engineering**

1. Selecting relevant features for enhancing model performance and prediction.
2. Create new meaningful features which will play a significant role on model performance.

* **Model Selection & Training**

Training different regression models namely Linear Regression, Ridge Regression, Lasso Regression, K- Nearest Neighbour, Decision Tree, Support Vector Regression (SVR), Random Forest, Extra Trees, AdaBoost, Gradient Boosting, XGBoost, Voting Regression, Stacking Regression.

* **Hyperparameter Tuning** to optimize model parameters for better accuracy.
* **Model Evaluation** using metrics like MAE and R² to assess model performance and select the best performing model.
* **Prediction & Deployment**

1. Using the best model to predict laptop prices
2. Creating a web application for users.

**DATA**

A glimpse of the dataset:



The dataset used for **Laptop Price Prediction** consists of **1,303 records**, each representing a laptop with multiple specifications. It contains **11 features** that influence the price, which is our target variable. Below is a breakdown of the key columns:

* **Company** – The brand or manufacturer of the laptop
* **TypeName** – The category of the laptop
* **Inches** – The display size of the laptop in inches
* **ScreenResolution** – The resolution of the display, affecting clarity and price
* **CPU** – The processor model and brand
* **RAM** – The memory size (e.g., 8GB, 16GB), impacting performance
* **Memory** – Storage capacity and type
* **GPU** – The graphics processing unit, important for gaming and design tasks
* **OpSys** – The operating system
* **Weight** – The weight of the laptop in kg, affecting portability
* **Price** – The target variable, representing the laptop's cost

We will apply various Feature Selection and Feature Engineering to obtain a final form of data which we will use for model training and testing.

The following Feature Engineering has been performed:

|  |  |
| --- | --- |
| Column | Operations |
| Unnamed:0 | Unnecessary column containing index. So, we can drop it. |
| Ram | Removing the text ‘GB’ and converting the datatype of the column into integers. |
| Weight | Removing the text ‘Kg’ and converting the datatype of the column into float. |
| Screen Resolution | From the data of this column, we can define new columns like ‘Touchscreen’, ‘IPS’, ‘PPI’. Then we can drop this ‘Screen Resolution’ column. |
| Inches | The importance of this column is fulfilled by the column ‘PPI’. So, we can drop it. |
| Company | No operation needed. |
| Type Name | No operation needed. |
| CPU | Only the text describing the CPU brand has been extracted and defined as a new column ‘CPU brand’ containing 5 different categories. Column ‘CPU’ has been dropped. |
| Memory | From the data of this column, we can define new columns like ‘HDD’, ‘SSD’, ‘Hybrid’, ‘Flash Storage’ which are different forms of memory and then we drop this ‘Memory’ column. We also drop ‘Hybrid’ and ‘Flash Storage’ column as they have very less correlation with price. |
| GPU | Only the text describing the GPU brand has been extracted and we define a new column ‘GPU brand’ containing 4 different categories. Column ‘GPU’ has been dropped. |
| OpSys | A new column ‘OS’ has been defined containing 3 categories and the column ‘OpSys’ has been dropped. |

Note:

* IPS (In-Plane Switching) is an electronic screen display technology used in LCDs (liquid crystal displays).
* Pixels per inch (PPI), or pixel density, on a laptop screen refers to the number of pixels packed into each square inch of the display, and generally, higher PPI means a sharper, more detailed image.

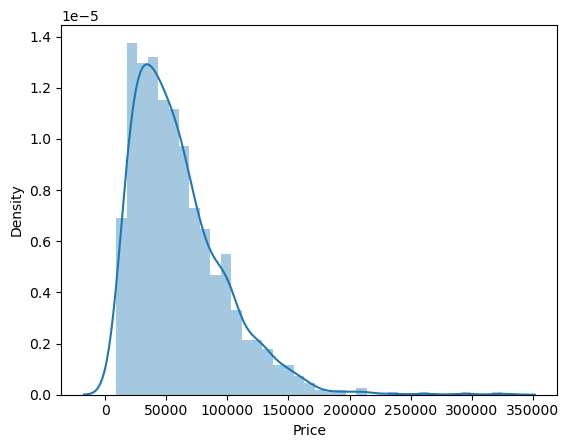
PPI = √ (horizontal pixels² + vertical pixels²) / (diagonal screen size in inches)

Ater performing the Feature Selection and Engineering,

* The columns ‘Company’, ‘Type Name’, ‘CPU brand’, ‘GPU brand’, ‘OS’ are categorical. So, we encode those columns with the help of One Hot Encoding to pass into the different ML models.
* The other columns like ‘RAM’, ‘Weight’, ‘Touchscreen’, ‘IPS’, ‘HDD’, ‘SSD’, ‘PPI’ are numerical. So, they do not require any kind of encoding.
* There are no missing values in the dataset. So we don’t have to handle any.

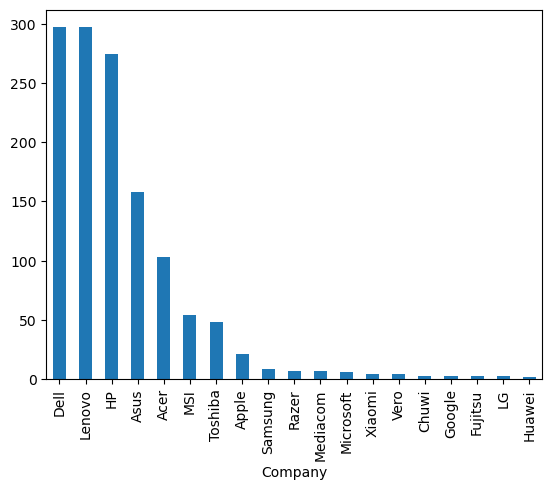
**RESULTS AND DISCUSSION**

* **EXPLORATORY DATA ANALYSIS (EDA)**
* **Distribution of Price**

****

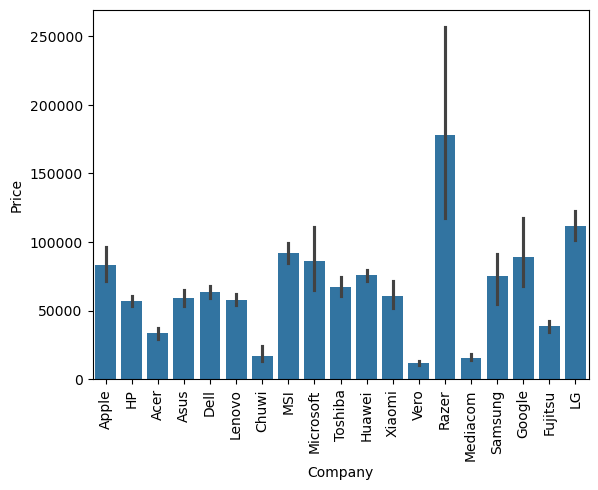
The distribution of price is skewed to the left.

* **Laptop sales of each company**

****

People buy laptops of companies like Dell, Lenovo, HP the most.

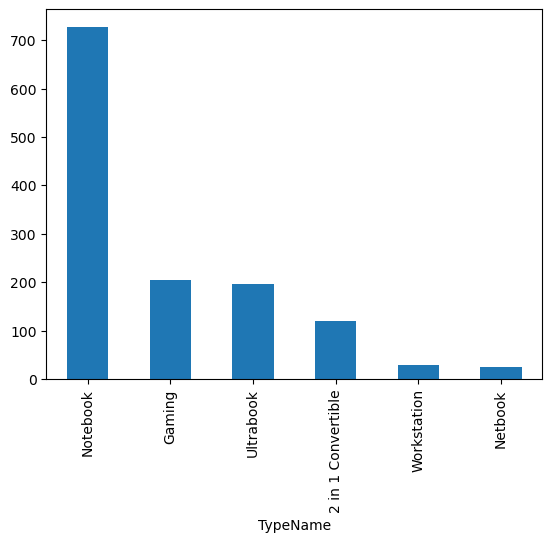
* **Average price of laptops of each company**

****

On an average, the laptops of ‘Razer’ are the most expensive.

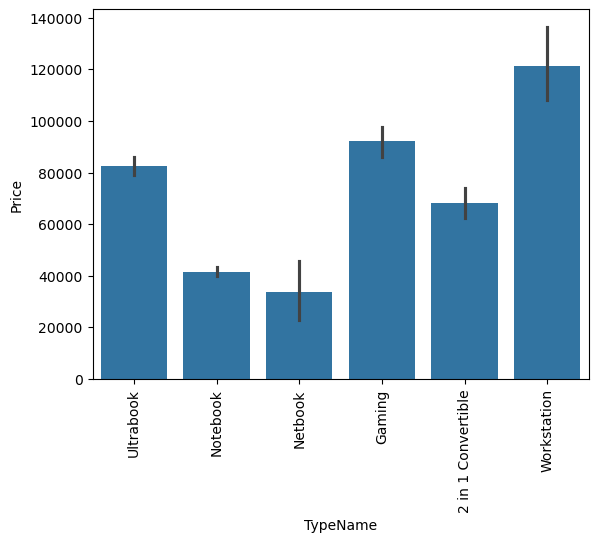
Meanwhile, companies like Dell, Lenovo, HP have relatively much lesser average price which makes them more affordable to buyers.

* **Number of laptops sold based on Type**

****

‘Notebooks’ are highly preferred by buyers.

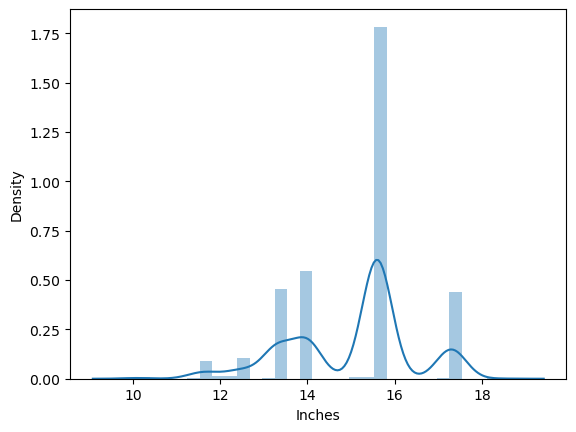
* **Average price based on Type**

****

On an average, ‘Workstations’ are the most expensive.

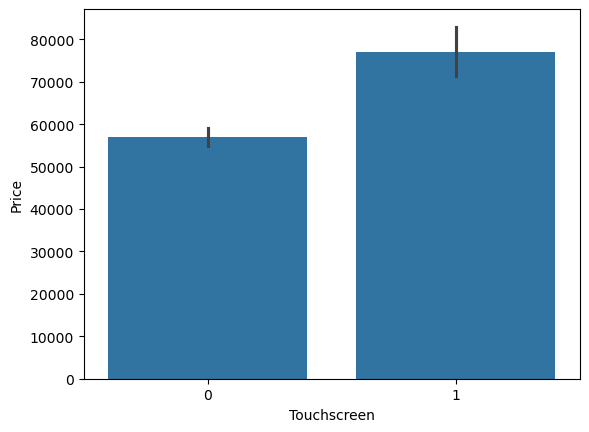
Notebook have relatively much lesser average price which makes them more affordable to buyers.

* **Distribution of Screen Size**

****

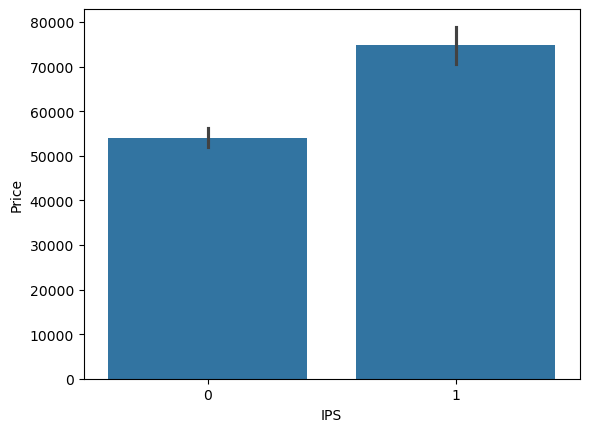
People buy laptops having screen size 15.6 inches the most.

* **Average price based on the presence or absence of the Touchscreen feature**

****

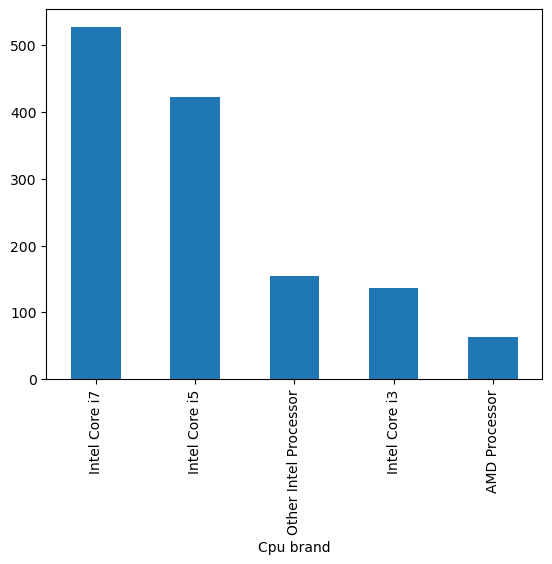
On an average, laptops having the touchscreen feature are relatively more expensive.

* **Average price based on the presence or absence of the IPS feature**

****

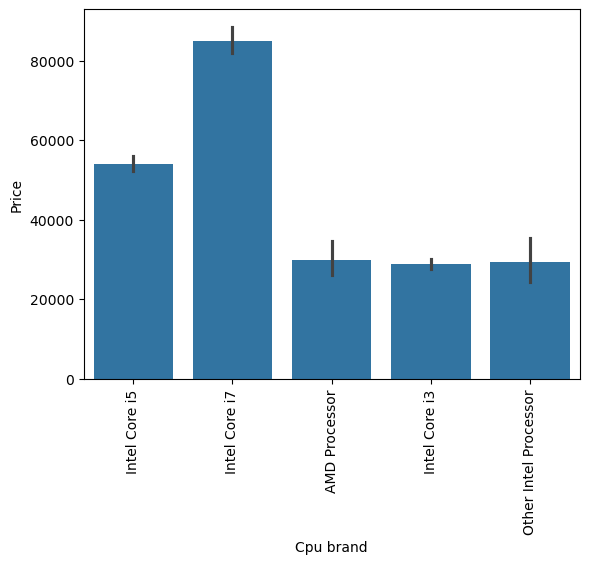
On an average, laptops having the IPS feature are relatively more expensive.

* **Numbers of laptops sold based on CPU brand**

****

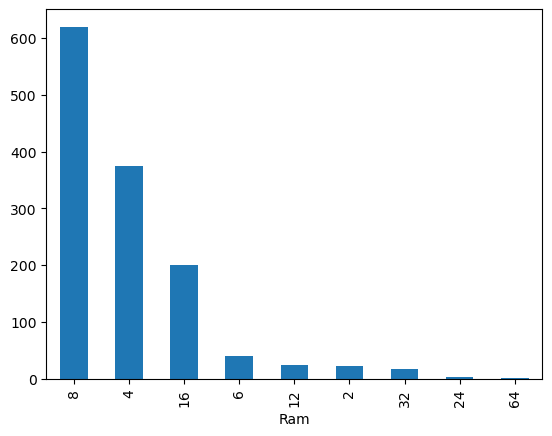
People buy laptops having CPU brand as ‘Intel Core i7’ and ‘Intel Core i5’ the most.

* **Average price of laptops based on CPU brand**

****

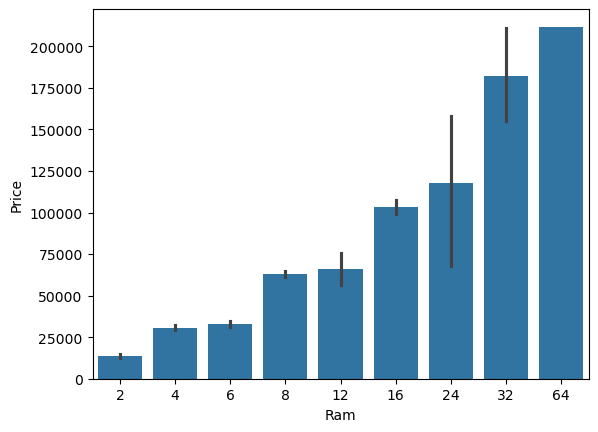
On an average, laptops having ‘Intel Core i7’ as their CPU brand are the most expensive

* **Number of laptops sold based on RAM**

****

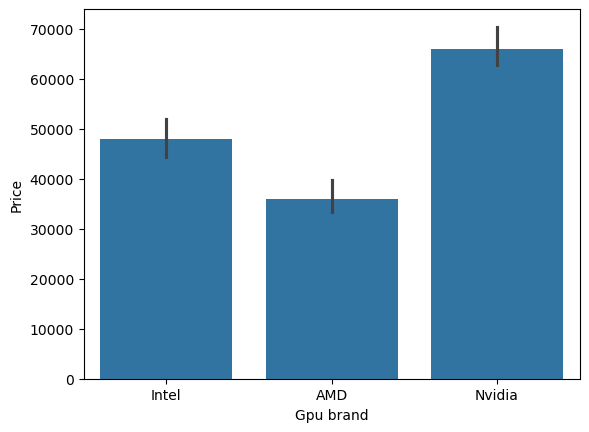
People buy 8GB RAM laptops the most.

* **Average price based on RAM**

****

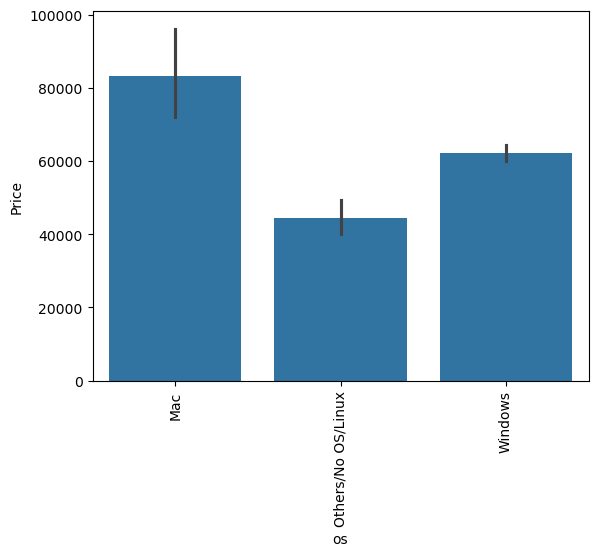
As the RAM size of a laptop increases, the price of it also increases indicating a linear positive relationship.

* **Average price based on GPU brand**

****

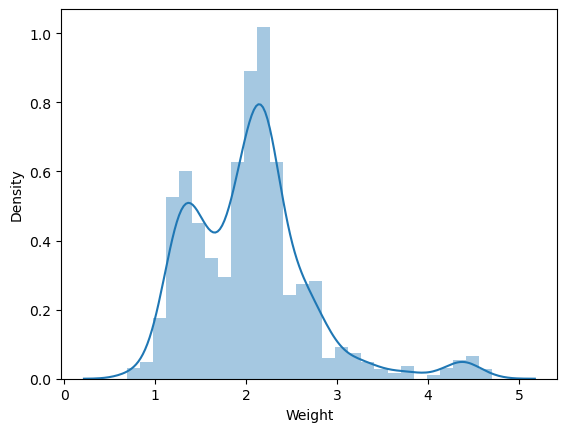
Laptops having GPU brand as ‘Nvidia’ are comparatively more expensive.

* **Average price based on Operating System**

****

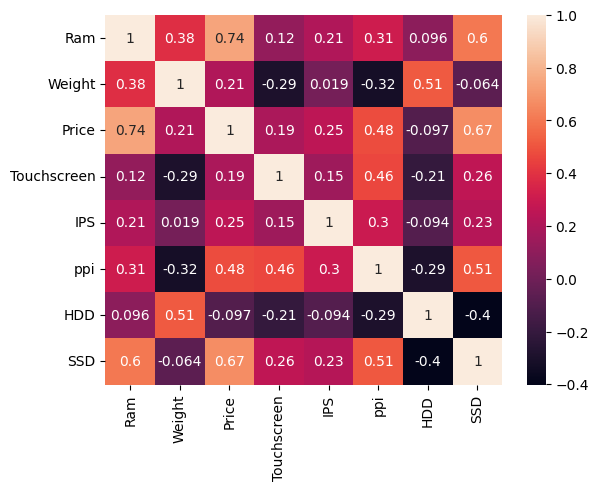
On an average, laptops having ‘Mac’ as the operating system are the most expensive.

* **Distribution of Weight**

****

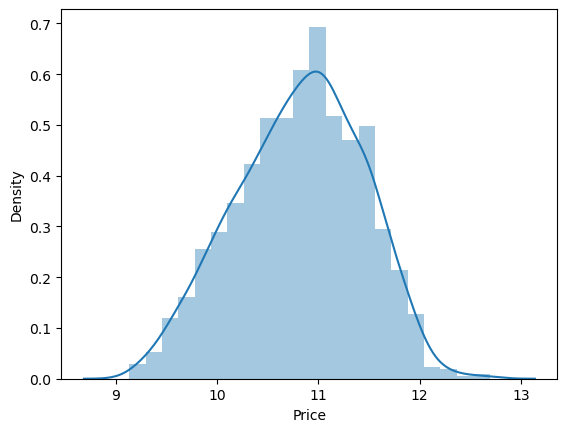
Weight of most of the laptops are distributed within 1.5 to 2.5 Kg.

* **Heatmap to understand the correlation between the numeric features**

****

1. Ram-Price, Ram-SSD, SSD-Price have strong positive relationship.
2. Weight-HDD, PPI-Price, Touchscreen-PPI, PPI-SSD have moderate positive relationship.
3. HDD-SSD has a moderate negative relationship.

* **Distribution of Price after taking logarithm of the prices**

****

The distribution of prices is left skewed. This skewness may influence the prediction. So, logarithm of the prices has been taken to remove the skewness.

* **COMPARISON OF MODEL PERFORMANCE**

|  |  |  |
| --- | --- | --- |
| **ML ALGORITHM** | **R- SQUARED** | **MEAN ABSOLUTE ERROR** |
| LINEAR REGRESSION | 0.8041 | 0.2090 |
| RIDGE REGRESSION | 0.8037 | 0.2102 |
| LASSO REGRESSION | 0.8041 | 0.2091 |
| KNN | 0.7978 | 0.2022 |
| DECISION TREES | 0.8376 | 0.1824 |
| SVR | 0.8215 | 0.1889 |
| RANDOM FOREST | 0.8839 | 0.1551 |
| EXTRA TREES | 0.8736 | 0.1591 |
| ADABOOST | 0.8728 | 0.1665 |
| GRADIENT BOOSTING | 0.9019 | 0.1402 |
| XGBOOST | 0.8958 | 0.1412 |
| VOTING REGRESSOR | 0.9040 | 0.1373 |
| STACKING REGRESSOR | 0.8979 | 0.1453 |

**Gradient Boosting Regressor** turned out to be the best model for this data and problem statement as it has the maximum R-Squared value and minimum Mean Absolute Error (MAE) among all the ML Algorithms considered.

**CONCLUSION**

* **Gradient Boosting Regressor** has been selected as◊◊◊◊◊◊◊ the best model for this data and problem statement as it has the maximum R-Squared value and minimum Mean Absolute Error (MAE) among all the ML Algorithms considered.

We are not selecting Voting Regressor because the improvement in R-squared as compared to Gradient Boosting is less and we will not select an ensemble method unless it results in a good improvement.

* To create the web app, we must store the pipeline corresponding to the Gradient Boosting model and the data in a pickle file.
* We then create our web app skeleton or the user interface using PyCharm.
* In our web app, the selected Gradient Boosting ML algorithm will be working to make predictions according to the specifications of the laptop given by the user through the user interface. The user will be provided with an estimated price of the laptop.