LaptopLens: Leveraging Machine Learning for Predictive Modeling of Laptop Prices

Analyzing Key Factors Influencing Laptop Pricing Through Advanced Algorithms

Seeta[1], SonaliYadav[2], Swastika Tomar[3], Swati Priyambda[4]

[1] Department of Artificial Intelligence and Data Science, Indira Gandhi Delhi Technical University for Women, New Delhi, India, sahuseeta06@gmail.com

[2] Department of Artificial Intelligence and Data Science, Indira Gandhi Delhi Technical University for Women, New Delhi, India, sonaliyadav0617@gmail.com

[3] Department of Artificial Intelligence and Data Science, Indira Gandhi Delhi Technical University for Women, New Delhi, India, lfis10cswastikatomar@gmail.com

[4] Department of Artificial Intelligence and Data Science, Indira Gandhi Delhi Technical University for Women, New Delhi, India, priyambda03@gmail.com

***Abstract*— Laptops have grown into necessary gadgets for both personal and business usage in the present-day rapidly evolving technological society. Predicting laptop costs is more crucial than ever because of the increasing demand for these devices due to digital education, remote work, and the necessity for high-performance machines.**

This study presents a supervised machine learning-based method for user-provided features, including CPU, RAM, storage, GPU, brand, screen resolution, and other essential aspects for estimating laptop prices. The model was developed using a dataset of 1,303 laptop instances with 12 attributes. Using two standard measures, such as mean absolute error (MAE) and R2 score, the process involved training machine learning models, including XG Boost, random forest, and linear regression, and evaluating their performance. With an R2 of 0.88, the random forest model's MAE is 0.15. The best fit for the predictive model is demonstrated by the random forest model's ability to discover the shortest MAE and the best R2 value that is closest to 1. With an 88% model efficiency, this indicates that the Random Forest model offers best accuracy.

In order to improve user experience, this model will be implemented on a web-based application using Flask, which will make it easier for customers to estimate laptop pricing according to their individual requirements and preferences.

Keywords—Laptop price, Linear Regression, Machine Learning, Random Forest, XG Boost Regressor, RAM, CPU, GPU.

I. INTRODUCTION

In the current age, a customer can find endless types of features in a single product, and this also applies while purchasing a laptop. This is why customers continually struggle with this challenge of obtaining an optimum value for their investment in a laptop while considering its plethora of features as well as their budgets. And the valuation of the global laptop market is increasing day by day. According to the statistics, the market was estimated to be worth USD 140.83 billion in 2022, and it is projected to increase at a compound annual growth rate of around 4.7% from 2023 to 2028 [1].

Now the challenge is that the intricacies of evaluating and selecting a laptop that strikes the perfect balance between performance and cost-efficiency have grown increasingly complex [8]. Even when confronted with laptops featuring similar specifications, the pricing of these models can vary drastically, often defying conventional logic [4]. This intricate interplay between feature parity and price diversity perplexes buyers, concealing the underlying factors contributing to these financial discrepancies [5][6].

This is why we need a system where a person can find an approximate price or, moreover, a price range based on their preferred features. Now our project's objective is to create an easy-to-navigate interface that estimates laptop prices based on input laptop characteristics. Information like CPU, RAM, storage, GPU, brand, screen resolution, and more will be provided by users. Based on these requirements, we applied ML models to anticipate prices with maximum accuracy and precision [4]. The aim is to flourish a simple and effective application that makes the process of purchasing a laptop easier for customers by enabling them to estimate laptop prices quickly and simply [8].

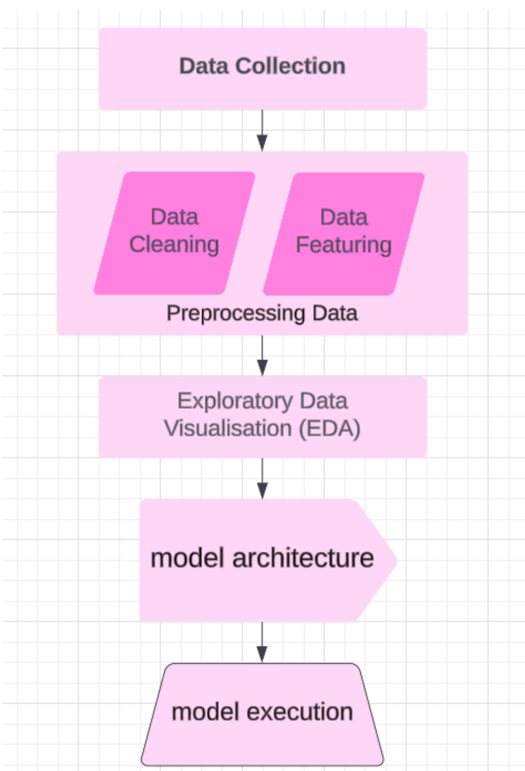
II. RELATED WORK

Using the collection of data gathered from the Kaggle website, Siburian et al. [4] employ the XGBoost, gradient boosting, and random forest algorithms for estimating laptop costs. The X G Boost algo regression yields a 92.77% R2 value.

The laptop pricing is determined by Shaik et al. [5] using a collection of data gathered from the Kaggle website using k-nearest neighbors (KNN), multiple linear regression, decision trees, and random forest Algo regressor. Random forest produced the best results, with an 88.75% R2 value and a mean absolute error of 0.1587.

III. RESEARCH METHODOLOGY

The below flow diagram shows the proposed method for the building of laptop price predictor model.



A. DATASET

The dataset utilized in this work, included 1303 samples, was then used for estimating laptop costs. With 12 characteristics covering diverse elements of laptops in various configurations, the set of data, officially titled "Laptop Price," mostly comes via the Kaggle platform. Hence, this data for laptop price prediction contains 12 columns and 1303 rows. By utilizing this dataset, the research can predict the prices of laptops with different configurations.

	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
5	Acer	Notebook	15.6	1366x768	AMD A9-Series 9420 3GHz	4GB	500GB HDD	AMD Radeon R5	Windows 10	2.1kg	21312.0000
6	Apple	Ultrabook	15.4	IPS Panel Retina Display 2880x1800	Intel Core i7 2.2GHz	16GB	256GB Flash Storage	Intel Iris Pro Graphics	Mac OS X	2.04kg	114017.6016
7	Apple	Ultrabook	13.3	1440x900	Intel Core i5 1.8GHz	8GB	256GB Flash Storage	Intel HD Graphics 6000	macOS	1.34kg	61735.5360
8	Asus	Ultrabook	14.0	Full HD 1920x1080	Intel Core i7 8550U 1.8GHz	16GB	512GB SSD	Nvidia GeForce MX150	Windows 10	1.3kg	79653.6000
9	Acer	Ultrabook	14.0	IPS Panel Full HD 1920x1080	Intel Core i5 8250U 1.6GHz	8GB	256GB SSD	Intel UHD Graphics 620	Windows 10	1.6kg	41025.6000
10	HP	Notebook	15.6	1366x768	Intel Core i5 7200U 2.5GHz	4GB	500GB HDD	Intel HD Graphics 620	No OS	1.86kg	20986.9920
11	HP	Notebook	15.6	Full HD 1920x1080	Intel Core i3 6006U 2GHz	4GB	500GB HDD	Intel HD Graphics 520	No OS	1.86kg	18381.0672
12	Apple	Ultrabook	15.4	IPS Panel Retina Display 2880x1800	Intel Core i7 2.8GHz	16GB	256GB SSD	AMD Radeon Pro 555	macOS	1.83kg	130001.6016
13	Dell	Notebook	15.6	Full HD 1920x1080	Intel Core i3 6006U 2GHz	4GB	256GB SSD	AMD Radeon R5 M430	Windows 10	2.2kg	26581.3920
14	Apple	Ultrabook	12.0	IPS Panel Retina Display 2304x1440	Intel Core M m3 1.2GHz	8GB	256GB SSD	Intel HD Graphics 615	macOS	0.92kg	67260.6720
15	Apple	Ultrabook	13.3	IPS Panel Retina Display 2560x1600	Intel Core i5 2.3GHz	8GB	256GB SSD	Intel Iris Plus Graphics 640	macOS	1.37kg	80908.3440
16	Dell	Notebook	15.6	Full HD 1920x1080	Intel Core i7 7500U 2.7GHz	8GB	256GB SSD	AMD Radeon R5 M430	Windows 10	2.2kg	39693.6000
17	Apple	Ultrabook	15.4	IPS Panel Retina Display 2880x1800	Intel Core i7 2.9GHz	16GB	512GB SSD	AMD Radeon Pro 560	macOS	1.83kg	152274.2400
18	Lenovo	Notebook	15.6	Full HD 1920x1080	Intel Core i3 7100U 2.4GHz	8GB	1TB HDD	Nvidia GeForce 940MX	No OS	2.2kg	26586.7200
19	Dell	Ultrabook	13.3	IPS Panel Full HD / Touchscreen 1920x1080	Intel Core i5 8250U 1.6GHz	8GB	128GB SSD	Intel UHD Graphics 620	Windows 10	1.22kg	52161.1200
20	Asus	Netbook	11.6	1366x768	Intel Atom x5-Z8350 1.44GHz	2GB	32GB Flash Storage	Intel HD Graphics 400	Windows 10	0.98kg	10224.4320
21	Lenovo	Gaming	15.6	IPS Panel Full HD 1920x1080	Intel Core i5 7300HQ 2.5GHz	8GB	128GB SSD + 1TB HDD	Nvidia GeForce GTX 1050	Windows 10	2.5kg	53226.7200
22	HP	Notebook	15.6	1366x768	AMD E-Series E2-9000e 1.5GHz	4GB	500GB HDD	AMD Radeon R2	No OS	1.86kg	13746.2400
23	Dell	2 in 1 Convertible	13.3	Full HD / Touchscreen 1920x1080	Intel Core i5 8250U 1.6GHz	8GB	256GB SSD	Intel UHD Graphics 620	Windows 10	1.62kg	43636.3200

B. DATA PREPROCESSING

One of the most crucial processes that improves the machine learning model's accuracy and efficiency is data preprocessing, which includes actions like normalization, noise removal, and checking for empty values.

1.Data Cleaning

In this step, the features of the dataset that can create various complications that will hinder the efficiency of the model need to be removed. So that model can operate without any bias, and it will not run into errors frequently. According to our dataset, this can include removing the duplicate rows, handling the missing values, and dropping the unnecessary column, such as Unnamed: 0. Moreover, there are rows in the data set consisting of alphanumeric values that cannot be interpreted by the machine learning model, as they are required to have numeric data to do the analysis. This is why we converted the weight replacement as an empty string rather than kg, an empty string for RAM rather than GB, and converted RAM into int32 format and weight into float32 type.

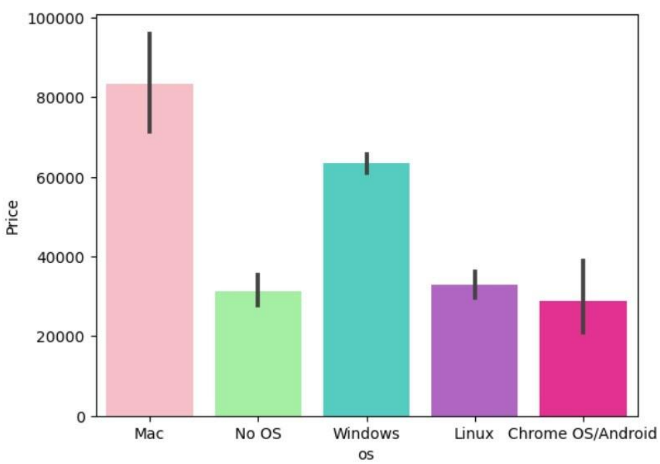
2.Feature Engineering

To help us conquer the challenge of doing the analysis efficiently, new and significant features can be extracted from some of the data set's attributes. We have modified the screen resolution feature in our dataset in various steps, which includes the addition of new features like touchscreen and IPS, which directly affect the prices of laptops. Similarly, from the CPU feature in the dataset, we have created a CPU brand feature that broadly identifies the brands, like Intel i5, i7, AMD, or Ryzen, as these are also the ones that highly affect the prices of laptops.

IV. EXPLORATORY DATA ANALYSIS (EDA)

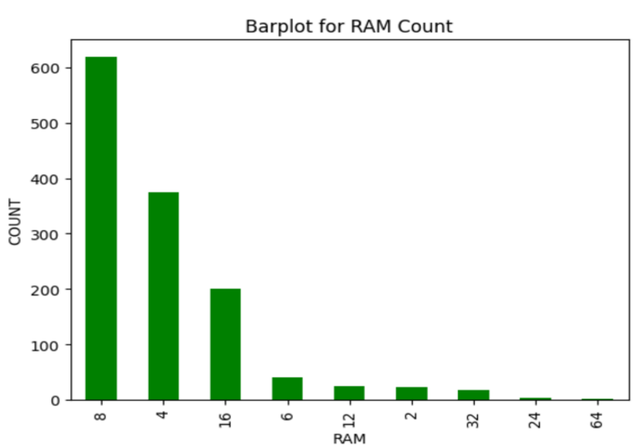
The process involves assessing gathered data to emphasize its essential characteristics, often utilizing visual tools for clarity. It involves techniques like plotting, calculating statistics, and identifying patterns or anomalies. Now that we have feature-engineered the given dataset, we can generate different visualizations, including graphs and tables, which will help us in analyzing the relationship between each feature and the variability in laptop prices. We have used the bar plot method from the matplotlib libraries to carry out the data visualization task. In seaborne library, we are able to test and validate the hypothesis or preliminary view regarding the impact of a feature on laptop pricing.

A. *OPERATING SYSTEM V/S PRICE*



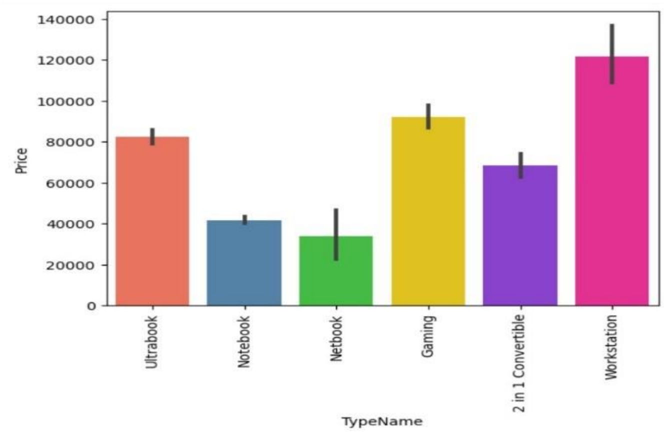
As shown in the above fig, we can conclude that operating systems like Mac are more expensive as compared to the ones that have Windows or any other operating system like Linux, chrome/android. Since we have laptops with no operating system, they come in less price as compared to the other ones.

B. *RAM V/S COUNT*



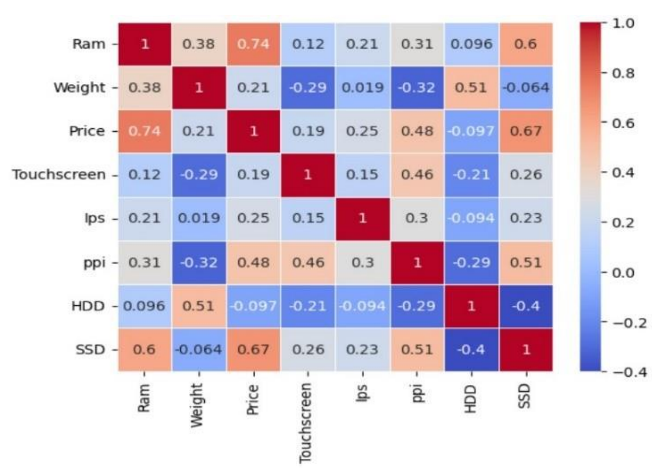
From this figure, it can be noted that the RAM of storage like 64 GB is rarer as compared to the RAM storage less than this. Consumers will also select 64G RAM capacity based on their requirement especially when they must handle large jobs or run numerous applications simultaneously.

C. *TYPENAME V/S PRICE*



We can infer from the above bar plot that workstation and gaming laptops are often more expensive than other laptop models. This is to be expected since these laptops frequently feature higher specifications (more memory, a faster CPU, etc.) to satisfy the needs of customers in a business setting.

D. *THE CORRELATION BETWEEN FEATURES*



The below figure is also known as a heatmap, and it is widely used in data visualization for the task of correlation between different features like RAM, weight, price, SSD, touchscreen, etc. This heatmap of a correlation matrix helps identify strong positive or negative relationships between variables. It displays data in a matrix format where individual values are represented by varying shades of color.

IV. MODEL BUILDING

The first step is that we had to split our data into the training and testing of about 15% testing and 85% training using sklearn test train split for validating the model with the right search queries. Therefore, we have split the training data as follows: (1106, 12) This explains there are going to be 1106 rows of data with all the columns included. Similarly, we have split the testing data in the manner of (197, 12); this is why there are going to be 326 rows of data with all the columns included. Now we are going to use various machine learning models that are present inside the sklearn module, and their performance is compared using metrics such as MAE and R2.

1.Prediction Algorithm Used

We have employed three various machine learning techniques that forecast the range of laptop prices. These include linear regression, the Random Forest, and XG Boost.

*A. **LINEAR REGRESSION ALGORITHM***

This method for predicting the results of a machine learning model is one of the fundamental statistical methods used to establish the connection among one or more variables that are independent and a variable that is dependent. It assumes that the relationship between these variables can be approximated with a straight line. Since here the target variable (laptop price) is continuous, which is ideal for regression algorithms like linear regression. It provides coefficients for each feature, making it easy to understand how each feature affects the price, as the positive coefficient will mean that the feature increases the prices and vice versa. It is highly useful for identifying the key predictors that are present in the given data as features.
$y=\beta 0+\beta 1X$

The variable that is dependent in the equation above is represented by y, and the variable that is independent is represented by x.. While β0 is the intercept or constant value and β1 is the slope of the line.

*B. **RANDOM FOREST ALGORITHM***

This approach is applicable for both classification and regression tasks. The basic idea is to create multiple decision trees during the training phase of the model and combine their outputs to mitigate the likelihood of overfitting and achieve more consistent predictive outcomes. Each decision tree in the forest is trained on a random subset of the data (with replacement) as part of the bagging process. This helps in introducing the diversity among the trees. Price prediction often involves complex, nonlinear relationships between features (e.g., RAM, processor type, PPI) and the target variable (price). Random Forest, as an ensemble of decision trees, captures these nonlinear interactions effectively. Also, some features like weight do not show much of an impact on the price of laptops as compared to the compared to the features like CPU brand and PPI. This method can easily handle feature importance as well as provide robustness to the outliers and noise.

*C. **XG BOOST ALGORITHM***

XGBoost is an advanced machine learning algorithm that leverages the gradient boosting framework, making it effective for building robust predictive models through the integration of multiple decision tree models. The key features that set this algorithm apart from others include regularization, handling missing values, parallelization, speed, and feature importance . Additionally, its features—such as parallel processing and feature importance evaluation—allow it to manage big datasets and avoid overfitting. Since our model consists of a very complex structured data set, this method helps a lot, as it is known for high performance and often outperforms other machine learning algorithms. Also, the laptop price prediction model often consists of complex interactions between multiple features, such as processor type, RAM, screen resolution, and storage type. Robustness to overfitting is one of the advantages.

2. Metrics that the model uses

In this paper, there are two typical metrics called MAE and R2 that are employed. The R2 value, also known as the coefficient of determination, is used to measure how well the real data and the model's predictions match. This calculates the target's or the dependent variable's percentage variance, which can be predicted based on the features, or independent variables. These measures have an interval of 0 to 1. Where the prediction matches the data accurately, the R-square is equal to 1and the actual value and the estimate of its value are similar. However, since the model fails to forecast any variation, we obtain an R-squared of 0 and it doesn't understand connection among variables that are dependent and independent.

$$SS_{total} = \sum_{i=1}^{n}(y_i - \bar{y})\hat{}2$$

$$SS_{residual} = \sum_{i=1}^{n}\left(y_i - h(x_i)\right)^2$$

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

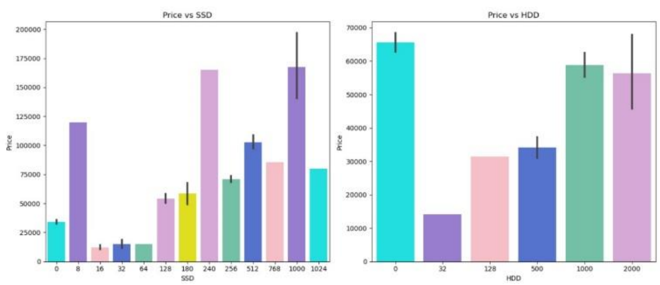
Now coming to MAE (Mean Absolute Error), which determines the average magnitude of errors in the estimations, without taking into account their trajectory (positive or negative). The characteristics of MAE include that this is a scale-independent measure that is sensitive to large errors but less than mean squared error (MSE). A lower MAE value indicates better model performance.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y_i} |$$

The key differences in the R^2 and MAE metrics include that $\mathbf{R^2}$ gives a relative measure of how well the model fits compared to a baseline (mean of the target variable) and **MAE** gives an absolute measure of average prediction error, useful for understanding the scale of errors.

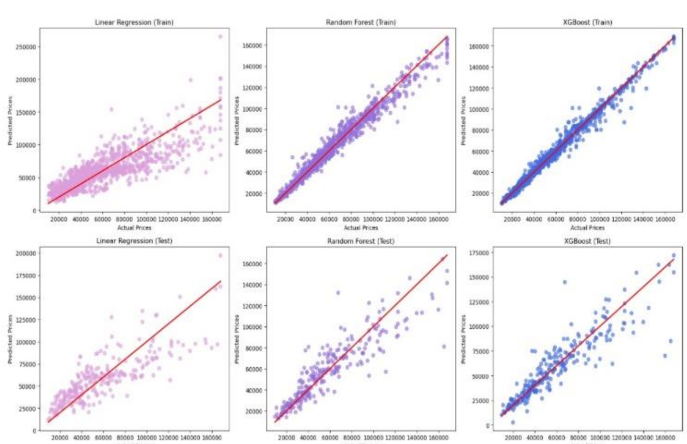
V. RESULTS AND DISCUSSION

A. ANALYSIS OF LAPTOP FEATURES



A laptop's HDD and SSD have a major effect on its expense because they determine the kind of storage, speed, and functionality. Although HDDs are relatively inexpensive compared to the SSDs, mainly because they use older technology like mechanical spinning disks as compared to the later ones. HDDs provide large storage capacity at a lower price, and hence the laptops having them tend to be more affordable. Now talking about SSDs, it uses flash memory, which is more costly to manufacture but offers better performance. SSDs drastically improve laptop speed, reducing boot times, file transfers, and application loading. This is why high-capacity SSDs significantly increase the price of laptops. They are commonly found in ultrabooks, gaming laptops, and professional laptops and are known for their durability, as they do not consist of any moving parts, which makes them less prone to physical damage. The above features are the ones that appeal to professionals and frequent travelers.

B. MODEL PERFORMANCE

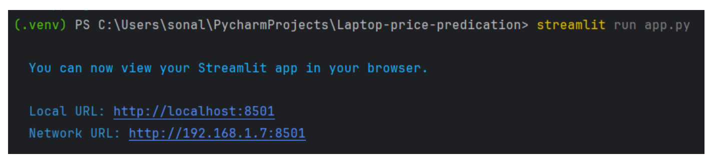


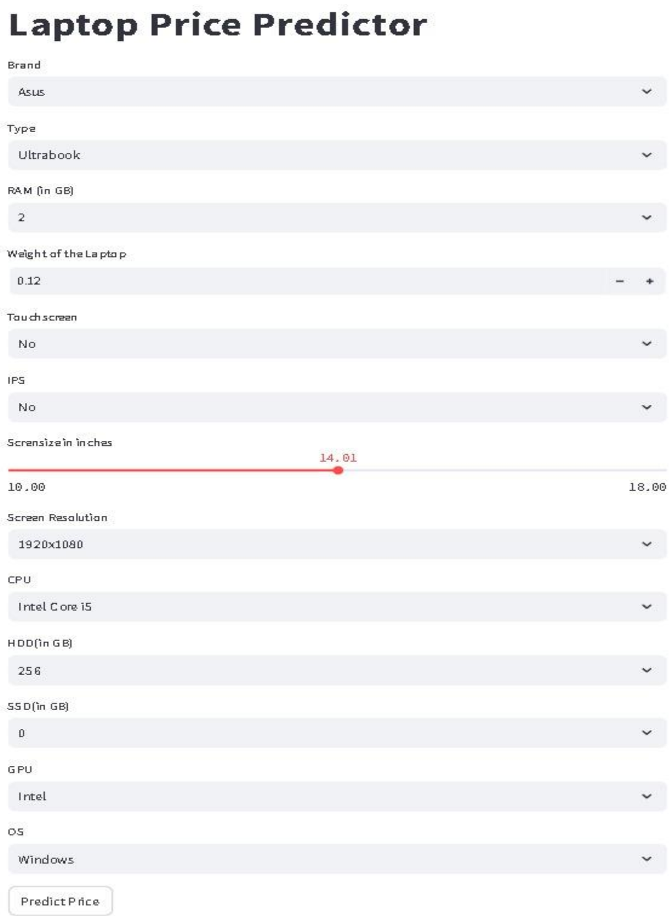
The scatter diagrams of the information regarding the XG Boost, random forest, and linear regression algorithms are presented above. To ensure the database's variety and security, the initial sample data was divided into a training collection of 1106 points of data and an evaluation data of 197 points of data..The mean absolute error (MAE) and R2 for the linear regression approach are 0.21 and 0.80, however The model based on random forests has an R2 of 0.88 and a mean absolute error of 0.15. Also, the mean absolute error is 0.16 and the R2 is 0.87 according to the XG boost approach.

It is clear when analyzing the regression model's measures MAE and R2 values for the model of random forest, linear regression, and XG Boost approaches, the model using random forests offers the best R2 statistic which is nearest to 1 and the least MAE. It means that the random forest approach matches prediction most effectively and provides the highest possible effectiveness. As a result, our model became 88% effective when estimating the cost of laptops depending upon the described features

VI. MODEL DEPLOYMENT

We have used Streamlit Library to build the user interface of our web application. It is an open-source Python framework that facilitates the development and dissemination of unique web apps for data science and machine learning. Since the project is still under development, and therefore we are working to add some new to this WebApp. The resulting interface of the same is shown below:





VII. CONCLUSION

While doing the project and data analysis, it can be concluded that popularity of laptop brands plays a significant role in price determination of the laptop. This study has accounted many parameters such as RAM, Company, Screen-Resolution, the Graphics Processing Unit (GPU), Operating System (OS), Memory (HDD/SSD) and Weight. For training and testing purposes, many mathematical models were imported from Scikit-learn like Linear Regression, Random Forest Regressor and XG Boost Regressor. Additionally, some tools, transformers and evaluation metrics like Column Transformer, Pipeline, One Hot Encoder, Mean Absolute Error, R2_score had been imported from the library. During model testing, it is found that the Random Forest approach has the best value for R2 (0.8856) and the least MAE (0.16009), showing the ultimate efficiency as well as feasibility to this forecast. By analyzing the future trends of laptop pricing, dealers can adjust prices to retain competition and attract customers. For customers, prediction of laptop prices is going to be beneficial in many ways. They can compare the predicted price of a laptop with its actual market price to evaluate whether they are getting a good deal or not. In addition, they can plan their budget-friendly deal, also educate customers in knowing what factors play a major role in laptop pricing and save their time in digital searching instead of doing manually

Nevertheless, there are significant limitations to the study, including the lack of enough assessments of how often factors affect laptop prices and the model's responsiveness to future shifts in market structures. More extended parameters and statistics are going to act as a basis for the models that predict in the following research.

VIII. REFERENCES

[1] “Global Laptop Market Report and Forecast 2024-2032,” Laptop Market Share, Size, Trends, Growth, Analysis 2024-2032, https://www.expertmarketresearch.com/reports/laptop-market (accessed Oct. 30, 2023).

[2] A. D. Siburian et al., “Laptop price prediction with machine learning using regression algorithm,” Jurnal Sistem Informasi dan Ilmu Komputer Prima(JUSIKOM PRIMA), vol. 6,no.1,pp.87–91,2022.

[3] M. A. Shaik, M. Varshith, S. SriVyshnavi, N. Sanjana and R. Sujith, "Laptop Price Prediction using Machine Learning Algorithms," 2022 International Conference on Emerging Trends in Engineering and Medical Sciences (ICETEMS), Nagpur, India, 2022, pp. 226-231, doi: 10.1109/ICETEMS56252.2022.10093357.

[4]https://www.kaggle.com/code/danielbethell/laptop-prices-prediction

[5]Siburian Astri Dahlia, Sitompul Daniel Ryan Hamonangan, Sinurat Stiven Hamonangan, et al. Laptop Price Prediction with Machine Learning Using Regression Algorithm. Jurnal Sistem Informasi dan Ilmu Komputer Prima (JUSIKOM PRIMA), 2022, 6(1): 87-91.

[6] Kushwaha Aaryan, Bansal Vasu, Goti Abuzar Shaikh, et al. Optimised Laptop Price Prediction. Journal of Computer Science Engineering and Software Testing, 2023, 9(1): 25-31.

[7] Priyama A, Abhijeeta RG, Ratheeb A, Srivastavab S. Comparative analysis of decision tree classification algorithms. International Journal of Current Engineering and Technology. 2013 Jun;3(2):334-7

[8] “Global Laptop Market Report and Forecast 2024-2032,” Laptop Market Share, Size, Trends, Growth, Analysis 2024-2032, https://www.expertmarketresearch.com/reports/laptop-market (accessed Oct. 30, 2023).

[9] M. A. Shaik, M. Varshith, S. S. Vyshnavi, N. M. S. Jana, and R. Sujith, ”Laptop Price Prediction using Machine Learning Algorithms,” 2022 International Conference on Emerging Trends in Engineering and Medical Sciences (ICETEMS), pp. 226-231, 2022.

[10] Zafar Jean-Denis, Stéphanie Himpens. Webscraping Laptop Prices to Estimate Hedonic Models and Extensions to Other Predictive Methods. 16th meeting of the Ottawa Group on Price Indices, Rio de Janeiro. 2019.