**CAPSTONE PRoject**

**LAPTOP PRICE PREDICTION USING MACHINE LEARNING**

SUBMISSION FOR PARTIAL FULFILLMENT OF REQUIREMENTS: PROFESSIONAL CERTIFICATE PROGRAM IN DATA SCIENCE FOR BUSINESS DECISIONS

BY

* SasiKumar
* Meera Palanivel
* Praveena Venu
* Niyati Mudliar
* Abhishu Tailor
* Padmavathy V
* Nithin Thomas

**GROUP6\_DSBD01**

INDIAN INSTITUTE OF MANAGEMENT KOZHIKODE

Date: 31 JULY 2024

**ACKNOWLEDGMENTS**

We would like to express our sincere gratitude to all those who have supported us throughout the completion of this Capstone Project.

First and foremost, we would like to thank our project advisor, Prof. Shovan Chowdhury, for his guidance, encouragement, and invaluable feedback throughout the project. His expertise and insights have been instrumental in shaping the direction and scope of this project.

We would also like to extend our appreciation to the faculty and staff of the Indian Institute of Management Kozhikode's Professional Certificate Program in Data Science for Business Decisions for providing us with the knowledge and skills necessary to undertake this project.

Furthermore, we would like to thank our family and friends for their unwavering support and encouragement throughout this journey. Their belief in us has been a constant source of motivation and inspiration.

Finally, we would like to acknowledge the contributions of each member of our group, SasiKumar, Meera Palanivel, Praveena Venu, Niyati Mudliar, Abhishu Tailor, Padmavathy V, and Nithin Thomas. Their dedication, hard work, and collaboration have been essential in the successful completion of this project.

Thank you all for your support and encouragement.

**EXECUTIVE SUMMARY**

The laptop market is a highly competitive and rapidly evolving industry. With the increasing demand for laptops, consumers are faced with a wide range of options, each with varying specifications and features. As a result, accurate and reliable price predictions are essential for both consumers and manufacturers to make informed decisions.

The capstone project undertaken by Group 6\_DSBD-01 as part of the Professional Certificate Program in Data Science for Business Decisions at the Indian Institute of Management Kozhikode aims to address this need. By developing a predictive model for estimating laptop prices based on main specifications and features, the project aims to provide valuable insights into the laptop market and help consumers make informed purchasing decisions.

The project's key objectives include performing descriptive analysis, engineering new features, and developing a predictive model using regression and testing techniques. The project's methodology involves data extraction, preprocessing, feature engineering, model development, and validation. The project's outcomes are expected to provide valuable insights into the laptop market and help consumers make informed purchasing decisions. The predictive model developed through this project can be used by retailers and manufacturers to optimize pricing strategies and improve profitability.

Overall, the capstone project undertaken by Group 6\_DSBD-01 is a significant contribution to the field of data science and has the potential to revolutionize the laptop market.

**TABLE OF CONTENT**

* Title Page
* Project Title
* Group Members
* Institution and Date
* Acknowledgments
* Executive Summary
* Table of Contents
* Main Content
* Introduction
* Background/Motivation
* Problem Identification and Industry Background
* Data Description
* Methodology
* Results/Analysis
  + EDA & Data Description
  + Predictive Models
* Interpretations/Discussion
* Conclusion
* References

**MAIN CONTENT**

**INTRODUCTION**

The laptop market is a highly competitive and rapidly evolving industry, with new models and features being introduced regularly. As consumers become increasingly reliant on technology, the demand for laptops has surged, leading to a wide range of options available in the market. However, with so many options available, it can be challenging for consumers to make informed decisions while purchasing laptops. This is where our project comes in.

Our project aims to analyze and predict the prices of laptops using machine learning techniques. By examining various specifications, pricing, and additional features, we intend to provide valuable insights that will empower consumers to make well-informed decisions while purchasing laptops. We will collect data from various sources, including online retailers, manufacturers, and consumer reviews, to create a comprehensive dataset that will be used to train our machine learning models.

We will also examine the impact of various factors such as processor speed, RAM, storage capacity, Graphic Card, Processor, and brand on laptop prices.

Through this project, we aim to create a valuable resource for consumers contemplating a laptop purchase. By presenting a detailed comparison of various models, including their technical specifications, cost, and additional features, we hope to simplify the decision-making process. Moreover, our analysis will shed light on the economic benefits of purchasing laptops, reinforcing the broader goal of promoting technological advancements. We will also provide insights into the environmental impact of laptop manufacturing and usage, highlighting the need for sustainable practices in the industry.

Ultimately, this project aspires to contribute to the growing body of knowledge on laptops and support the transition to a more technologically advanced future. By providing consumers with valuable insights and information, we hope to empower them to make informed decisions while purchasing laptops, leading to a more efficient and sustainable laptop market.

**BACKGROUND**

The laptop market has been growing rapidly over the past few years, with more and more people relying on laptops for work, education, and entertainment. With the increasing demand for laptops, there has been a surge in the number of laptop brands and models available in the market. This has made it challenging for customers to choose the right laptop that fits their budget and requirements. Moreover, the prices of laptops keep fluctuating due to various factors such as market demand, supply chain disruptions, and technological advancements. Therefore, there is a need for a reliable and accurate laptop price prediction system that can help customers make informed decisions.

**MOTIVATION**

The motivation behind the capstone project "Laptop Price Prediction using Machine Learning" is to develop a machine learning model that can predict the price of a laptop accurately. The model will consider various factors such as brand, processor, RAM, storage, graphics card, and other specifications to predict the price of a laptop. The project aims to help customers make informed decisions by providing them with accurate price predictions, which will save them time and money. Additionally, the project will also benefit laptop manufacturers and retailers by providing them with insights into the market trends and customer preferences. The project will involve collecting data on various laptop models and their prices from different sources such as online retailers, manufacturer websites, and market research reports. The data will be cleaned and preprocessed to remove any inconsistencies and errors. The preprocessed data will then be used to train and test the machine learning model. The machine learning model will be developed using various algorithms such as linear regression, decision trees, and random forests. The model will be trained on a subset of the data and tested on another subset to evaluate its accuracy. The model will be fine-tuned by adjusting the hyperparameters to improve its performance. Once the model is developed and tested, it will be deployed as a web application that customers can use to predict the price of a laptop. The web application will take input from the user on the laptop specifications and provide an accurate price prediction. The web application will also provide insights into the market trends and customer preferences, which will be useful for laptop manufacturers and retailers.

Overall, the project will contribute to the development of a more efficient and effective laptop market by providing customers with accurate price predictions and manufacturers and retailers with insights into the market trends and customer preferences.

**PROBLEM IDENTIFICATION AND INDUSTRY BACKGROUND**

**Problem Identification:**

The laptop market is highly competitive, with numerous brands and models available in the market. With the increasing demand for laptops, it has become challenging for customers to choose the right laptop that fits their budget and requirements. Moreover, the prices of laptops keep fluctuating, making it difficult for customers to predict the price of a laptop accurately. This leads to customers either overspending or underspending on laptops, which can result in dissatisfaction and financial losses. Additionally, laptop manufacturers and retailers face challenges in predicting the demand for different laptop models and setting the right prices, which can impact their profitability.

**Industry Background:**

The laptop market has been growing rapidly over the past few years, with more and more people relying on laptops for work, education, and entertainment. According to a report by Statista, the global laptop market size was valued at USD 109.6 billion in 2020 and is expected to reach USD 146.5 billion by 2027. The market is highly competitive, with numerous brands such as Dell, HP, Lenovo, Apple, and Asus competing for market share. The laptop market is segmented based on various factors such as brand, processor, RAM, storage, graphics card, and other specifications. Each segment has its own set of customers with different preferences and budgets.

Therefore, it is essential for laptop manufacturers and retailers to understand the market trends and customer preferences to develop and market the right laptop models. However, predicting the demand for different laptop models and setting the right prices is a challenging task due to various factors such as market demand, supply chain disruptions, and technological advancements.

Therefore, there is a need for a reliable and accurate laptop price prediction system that can help customers make informed decisions and provide manufacturers and retailers with insights into the market trends and customer preferences.

**Laptop Manufacturers:**

Laptop manufacturers play a crucial role in the laptop market. They are responsible for designing, developing, and manufacturing different laptop models that cater to the needs and preferences of customers. Some of the major laptop manufacturers in the market include Dell, HP, Lenovo, Apple, Asus, Acer, and MSI.

Each laptop manufacturer has its own strengths and weaknesses, and they compete with each other to gain market share. For instance, Dell is known for its high-quality business laptops, while Apple is known for its premium laptops with a sleek design. Lenovo, on the other hand, offers a wide range of laptops that cater to different customer segments, from budget laptops to high-end gaming laptops.

Laptop manufacturers face challenges in predicting the demand for different laptop models and setting the right prices. They need to understand the market trends and customer preferences to develop and market the right laptop models. Additionally, they need to keep up with the technological advancements and offer laptops with the latest features and specifications to stay competitive in the market.

Therefore, a reliable and accurate laptop price prediction system can benefit laptop manufacturers by providing them with insights into the market trends and customer preferences. This can help them develop and market the right laptop models and set the right prices to maximize their profitability.

### **DATA DESCRIPTION**

**Overview**

This dataset contains detailed information about various Laptop models. It includes technical specifications, performance metrics, and price details.

**Dataset Characteristics**

* The dataset contains information on **823 laptop** models.
* Each laptop has **19 features** including brand, processor, RAM, storage, and more.
* The goal is to **predict laptop prices** based on their specifications.

**Column Descriptions**

1. Brand (Int) - The manufacturer or brand of the Laptop.
2. Processor\_brand (Int) - The specific model name of the Laptop
3. Processor\_name (Int) - The name of the processor used by the specific laptop model
4. Processor\_gnrtn (Int) - Indicates the processor generation
5. Ram\_gb (Int) - The capacity of the RAM in GB
6. Ram\_type (Int) - Indicates the RAM Type
7. Ssd (Int) - The capacity of the SSD (Solid-State Drive) in GB
8. Hdd (Int) - The capacity of the HDD (Hard Disk Drive) in GB
9. Os (Int) - The name of the Operating System
10. Os\_bit (Int) - The type of Operating system in bit
11. Graphic\_card\_gb (Int) - The capacity of the graphic card in GB
12. Weight (Int) - The type of Laptop Model
13. Warranty (Int) - The duration of the warranty duration
14. Touchscreen (Int) - Touchscreen availability
15. Msoffice (Int) - Default Microsoft office suit availability
16. Price (Float) - The laptop model price in EURO
17. Rating (Int) - The start rating of the laptop model
18. Number of Ratings (Float) - The number of Ratings of each models
19. Number of Reviews (Float) - The number of reviews of each models

**METHODOLOGY**

**1. Data extraction:**

We extracted the data on Laptop models, specifications, pricing from <https://www.kaggle.com/> and <https://ev-database.org/> . Both the sites had data pertaining to Laptop models existing in CY2024.

**2. Data Preprocessing**

We performed pre-processing on the data, which involved tasks such as data cleaning, to make it more suitable for the analysis we intended to conduct. Additionally, we familiarized ourselves with the specifications of different laptop models and their significance.

**3. Exploratory data analysis & Descriptive Analysis:**

Data Visualization: The next step was to visualize the data using various plots such as scatter plots, histograms, and box plots. This helped in identifying any patterns or trends in the data.

Correlation Analysis: Correlation analysis was performed to identify any relationships between the different variables. This helped in understanding which variables had the most impact on laptop prices.

Basic structure of the database and the various variables was summarized using Python and the same was recorded in the Descriptive Analytics.

**4. Feature engineering:**

This involved selecting the most important variables that had the most impact on laptop prices. Some of the features that were extracted included processor speed, RAM, storage capacity, screen size, and brand. This involved converting categorical variables into numerical variables using techniques such as one-hot encoding or label encoding. Feature scaling was performed to ensure that all the features were on the same scale. This helped in improving the accuracy of the machine learning algorithms. Techniques such as standardization or normalization were used for feature scaling

**5. Model development:**

After the EDA and studying the database description summary, we proceeded for developing Linear Regression Model. And develop the Xg boost model performance using Python,

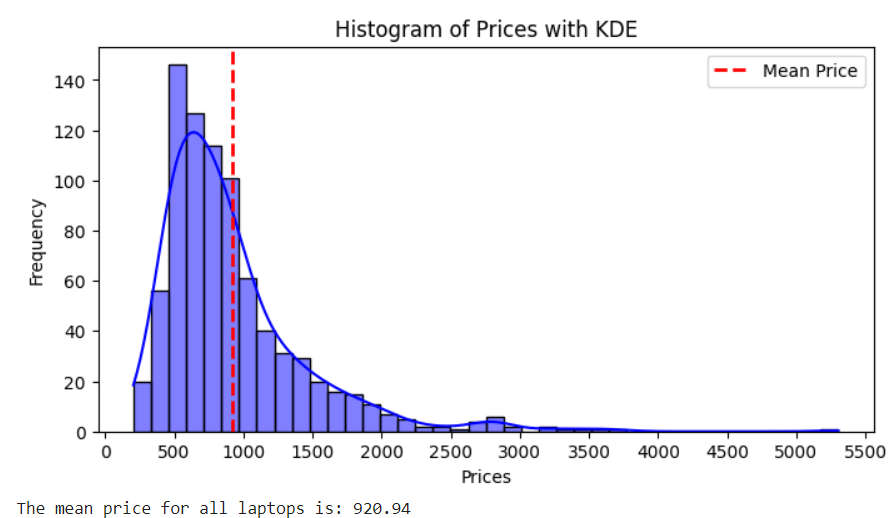
**6. Model Testing & Evaluation:**

We evaluated the linear regression & Xg Boost model outputs to select the best fit model. In doing so, we analysed the various Statistical parameters generated in the model output using Pyton. Key parameters included: R2 , adjusted R2, RSE.

**RESULTS/ANALYSIS**

**Exploratory Data Analysis (EDA)**

**Price Distribution**



The graph is a histogram with a kernel density estimate (KDE) overlay, displaying the distribution of prices. Here is a breakdown of its components:

1. **Histogram**: The bars represent the frequency of prices within specific intervals (bins). The height of each bar indicates how many observations fall within that price range. In this graph, most of the prices are clustered at the lower end, indicating that lower prices are more common.
2. **Kernel Density Estimate (KDE)**: The blue curve represents a smoothed version of the histogram, showing the estimated probability density function of the prices. This helps visualize the distribution's shape without relying solely on the bins' boundaries.
3. **Red Dashed Line**: This line marks the mean price, labelled as "Mean Price" in the legend. The placement of this line suggests that the mean price is around the value where the line intersects the x-axis.
4. **Right Skewness**: The distribution has a longer tail on the right side, indicating a right-skewed distribution. This means that while most prices are low, there are some higher-priced outliers that pull the mean to the right.

Overall, the graph shows that most prices are concentrated in the lower range, with a few instances of significantly higher prices, resulting in a right-skewed distribution. The KDE curve provides a smooth approximation of this distribution.

**Brand Price Comparison**

A chart with different colored boxes

Description automatically generated

This boxplot presents a comparison of product prices across different brands.

**Key Insights:**

* **Price Range:** The vertical axis represents price, indicating that Apple and MSI products generally have higher price points compared to other brands.
* **Distribution:** The boxplots show the distribution of prices for each brand. The box represents the interquartile range (IQR), containing 50% of the data, with the median price indicated by the line within the box.
* **Outliers:** The dots outside the whiskers represent outliers, suggesting that some products from certain brands have significantly higher prices compared to the rest.

Overall, the graph suggests that Apple and MSI products tend to be more expensive, while brands like Avita and Lenovo offer products at lower price points.

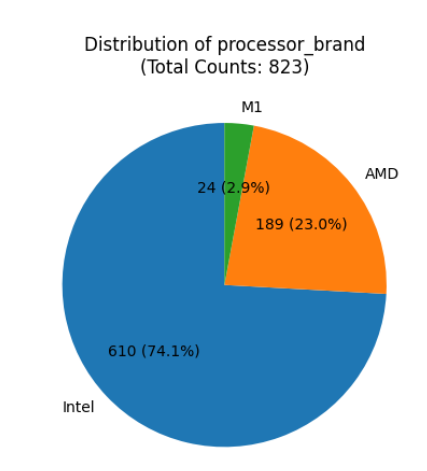
A graph of a number of colored bars

Description automatically generated with medium confidence The graph displays a bar chart representing the mean price for each brand. The height of each bar corresponds to the average price of products from that brand.

**Observations:**

1. **Price Variation:** There is a significant difference in mean prices across brands.
2. **Highest Mean Price:** APPLE has the highest average price among all brands, with the bar reaching the top of the chart.
3. **Lowest Mean Price:** DELL and Avita appear to have the lowest average prices, with their bars starting close to the bottom of the chart.
4. **Price Distribution:** The bars for acer, Lenovo, and HP have similar heights, indicating their mean prices are relatively close.

Overall, this graph provides a visual comparison of the average product prices for different brands. It highlights that APPLE products are significantly more expensive on average compared to other brands.

**Processor Brands**

The chart displays the distribution of processor brands, represented as percentages of the total count (823).

**Breakdown:**

* **Intel:** The largest slice, occupying 74.1% of the chart, indicating that 610 out of 823 processors are Intel.
* **AMD:** The second-largest slice, comprising 23.0% or 189 processors.
* **M1:** The smallest slice, representing 2.9% or 24 processors.

**Key Insights:**

* Intel processors are the most prevalent in the dataset, accounting for over three-quarters of the total.
* AMD processors hold a significant share (23%), while M1 processors are relatively less common.

**Additional Considerations:**

* The pie chart effectively visualizes the distribution of processor brands and the relative proportions of each category.
* However, for more precise data analysis, consider using numerical values or bar charts alongside the pie chart.

Overall, the chart provides a clear overview of the dominant processor brand (Intel) and the distribution of other brands (AMD and M1) within the dataset.

**Light Weight Categories**

A pie chart with numbers and a number of weight

Description automatically generated

The pie chart illustrates the distribution of weight across three categories: Casual, Thin & Light, and Gaming.

* **Casual:** The largest slice, occupying 63.2% of the chart, indicating that 520 out of 823 items fall into this category.
* **Thin & Light:** The second-largest slice, comprising 32.1% or 264 items.
* **Gaming:** The smallest slice, representing 4.7% or 39 items.

**Key Insights:**

* **Dominance of Casual:** The "Casual" category is the most prevalent, accounting for over two-thirds of the total weight distribution.
* **Moderate Thin & Light:** The "Thin & Light" category constitutes a significant portion, representing around one-third of the data.
* **Small Gaming Segment:** The "Gaming" category is relatively small, comprising less than 5% of the total weight.

Overall, the pie chart effectively visualizes the distribution of weight across the three categories, highlighting the dominance of the "Casual" category and the smaller proportions of "Thin & Light" and "Gaming".

**Touchscreen V/s non-touchscreen**

A blue and orange pie chart

Description automatically generated

The pie chart illustrates the proportion of devices with and without touchscreens within a dataset of 823 devices.

* **Yes:** The smaller slice, representing 11.8% or 97 devices, indicates the proportion of devices with touchscreens.
* **No:** The larger slice, comprising 88.2% or 726 devices, signifies the proportion of devices without touchscreens.

**Key Insights:**

* **Dominance of Non-Touchscreen Devices:** The majority (88.2%) of devices in the dataset do not have touchscreens.
* **Smaller Touchscreen Segment:** Only 11.8% of the devices are equipped with touchscreens.

Overall, the pie chart effectively visualizes the distribution of touchscreen devices versus non-touchscreen devices within the dataset, clearly demonstrating the predominance of non-touchscreen models.

**Operating System Distribution**

**A graph of a computer

Description automatically generated**

The chart displays the count of laptops categorized by operating system (OS) and bit version (32-bit or 64-bit).

**Breakdown:**

* **Windows:**
  + 64-bit: The tallest bar, indicating a significantly higher count of Windows laptops with 64-bit architecture.
  + 32-bit: A smaller bar, suggesting a considerable number of Windows laptops using 32-bit architecture.
* **DOS:**
  + 64-bit: No bar present, indicating no DOS laptops with 64-bit architecture.
  + 32-bit: A very small bar, suggesting a negligible count of DOS laptops with 32-bit architecture.
* **Mac:**
  + 64-bit: A small bar, indicating a relatively low count of Mac laptops with 64-bit architecture.
  + 32-bit: No bar present, indicating no Mac laptops with 32-bit architecture.

**Key Insights:**

* Windows Dominance: Windows is the most prevalent OS, with a significant number of both 64-bit and 32-bit versions.
* DOS Decline: DOS laptops are extremely rare, with only a few 32-bit systems present.
* Mac Presence: Mac laptops are less common than Windows, primarily using 64-bit architecture.

Overall, the chart effectively visualizes the distribution of laptops across different OS and bit versions, highlighting the dominance of Windows and the scarcity of DOS systems.

**A graph of different colored boxes

Description automatically generated**

The boxplot illustrates the distribution of prices for laptops categorized by their operating system (OS) and bit version (32-bit or 64-bit). Each box represents a group of laptops with the same OS.

**Boxplot Components:**

* **Median:** The horizontal line within each box represents the median price for that OS group.
* **Interquartile Range (IQR):** The box itself represents the IQR, containing 50% of the data points.
* **Whiskers:** The lines extending from the box indicate the range of data within 1.5 times the IQR. Data points beyond this range are considered outliers.
* **Outliers:** Individual data points plotted as circles or dots beyond the whiskers represent outliers, which are significantly higher-priced laptops compared to the rest of the group.

**Observations:**

* **Price Variation:** There is a significant price difference between OS groups and even within the same OS but different bit versions.
* **Windows:**
  + Both 32-bit and 64-bit Windows laptops have a wide price range, with several outliers on the higher end.
  + The median price for 64-bit Windows laptops is higher than that of 32-bit Windows laptops.
* **DOS:**
  + Only 32-bit DOS laptops are included in the data.
  + The price range for DOS laptops is relatively narrow compared to other OS groups.
* **Mac:**
  + Only 64-bit Mac laptops are included.
  + The median price for Mac laptops is significantly higher than both Windows and DOS laptops.
  + The price range is wide, with several high-priced outliers.

Overall, the boxplot effectively visualizes the distribution of laptop prices across different OS and bit versions, highlighting the price differences between OS groups and the presence of high-priced outliers in Windows and Mac categories.

**Storage Options**

**A purple bar graph with numbers

Description automatically generated**

The chart illustrates the frequency or count of different HDD sizes within a dataset.

* **Dominant Size:** The HDD size of "0 GB" has the highest count, represented by the tallest bar. This indicates that a significant number of devices in the dataset have no HDD.
* **Other Sizes:** The sizes "1024 GB" and "512 GB" have moderate counts, suggesting a fair number of devices with these HDD capacities. "2048 GB" has the lowest count, indicating fewer devices with this large HDD size.

**Key Insights:**

* Most devices in the dataset appear to lack an HDD (0 GB).
* There is a reasonable proportion of devices with 1024 GB and 512 GB HDDs.
* Large HDDs (2048 GB) are less common in the dataset.

**Possible Implications:**

The data suggests that most devices analysed might be low-end or entry-level devices without storage drives, or they could be devices with primary storage solutions like SSDs. The presence of 1024 GB and 512 GB HDDs indicates a mix of devices with varying storage needs.

**A graph with purple bars

Description automatically generated with medium confidence**

The chart illustrates the frequency or count of different SSD sizes within a dataset.

**Key Observations:**

1. **Dominant Size:** The SSD size of "512 GB" has the highest count, represented by the tallest bar. This indicates that a significant number of devices in the dataset have 512 GB SSDs.
2. **Other Sizes:**
   * "256 GB" has a notably high count, suggesting a considerable number of devices with this SSD capacity.
   * "0 GB" has a moderate count, indicating a fair number of devices might have no SSD or have other primary storage solutions.
   * "128 GB" has a lower count, suggesting fewer devices with this SSD size.
   * "1024 GB" and larger sizes have very low counts, indicating their rarity in the dataset.

Overall, the chart provides a clear visual representation of SSD size distribution within the dataset. Many devices seem to have either 512 GB or 256 GB SSDs, with fewer devices having smaller or larger capacities.

**RAM Configuration**

A purple bar graph with numbers

Description automatically generated

The chart illustrates the frequency or count of different RAM sizes within a dataset.

**Key Observations:**

* **Dominant Size:** The RAM size of "8 GB" has the highest count, represented by the tallest bar. This indicates that a significant number of devices in the dataset have 8 GB RAM.
* **Other Sizes:**
  + "4 GB" has a notably high count, suggesting a considerable number of devices with this RAM capacity.
  + "16 GB" has a lower count than 8 GB but is still significant, indicating a fair number of devices with this RAM size.
  + "32 GB" has the lowest count, suggesting fewer devices with this RAM capacity.

Overall, the chart provides a clear visual representation of RAM size distribution within the dataset. Many devices seem to have either 8 GB or 4 GB RAM, with a smaller proportion having 16 GB, and very few with 32 GB.

**Possible Implications:**

The data indicates that most devices analysed are likely mid-range, with sufficient RAM for general tasks. A notable presence of devices with 4 GB RAM suggests some older or budget-oriented models. The smaller proportion of devices with 16 GB and 32 GB RAM points to these configurations being more common in high-performance or specialized devices.

**Processor Brand Price Impact**

**A graph of a box plot

Description automatically generated**

**Observations:**

* **Price Variation:** There is a significant price difference between processor brands.
* **Intel and AMD:** Both Intel and AMD have a wide price range, with several outliers on the higher end. The median price for Intel is slightly lower than AMD.
* **M1:** M1 processors have a higher overall price range compared to Intel and AMD. The median price for M1 is significantly higher than both Intel and AMD.

Overall, the box plot effectively visualizes the distribution of laptop prices across different processor brands, highlighting the price differences between brands and the presence of high-priced outliers in all categories.

**Additional Insights:**

* M1 processors tend to be more expensive than Intel and AMD processors.
* There are some very high-priced laptops with Intel and AMD processors, indicating potential premium models or configurations.

**Processor Brand Price Impact**

**A graph of a box plot

Description automatically generated**

**Observations:**

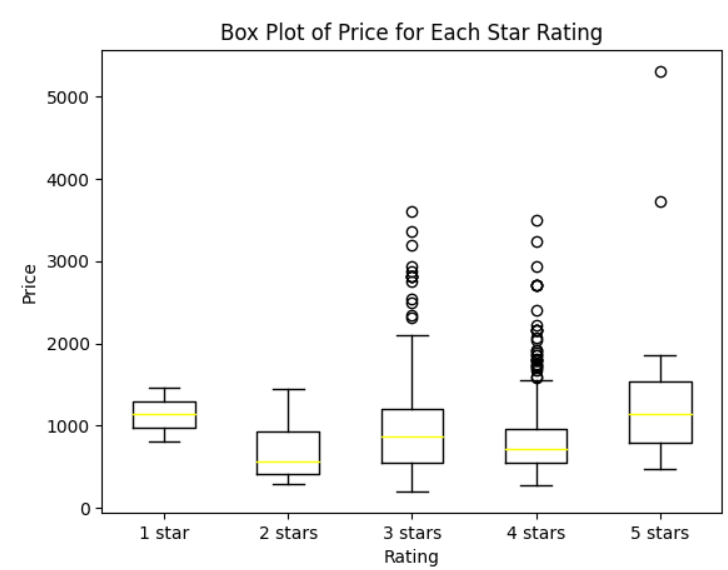
* **Price Variation:** There is a significant price difference between processor brands.
* **Intel and AMD:** Both Intel and AMD have a wide price range, with several outliers on the higher end. The median price for Intel is slightly lower than AMD.
* **M1:** M1 processors have a higher overall price range compared to Intel and AMD. The median price for M1 is significantly higher than both Intel and AMD.

Overall, the box plot effectively visualizes the distribution of laptop prices across different processor brands, highlighting the price differences between brands and the presence of high-priced outliers in all categories.

**Additional Insights:**

* M1 processors tend to be more expensive than Intel and AMD processors.
* There are some very high-priced laptops with Intel and AMD processors, indicating potential premium models or configurations.

**Price By Star Rating**



**Key Observations**

* **Price Distribution:** The price distribution varies across different star ratings.
* **Outliers:** There are several outliers, especially in the higher star rating categories, indicating products with significantly higher prices compared to others in the same category.
* **Median Price:** The median price tends to increase with higher star ratings, suggesting that higher-rated products generally have higher prices.
* **Price Range:** The range of prices within each star rating category is quite wide, indicating a diverse price spectrum for products at each rating level.

**Overall Interpretation**

The chart suggests a correlation between product price and star rating. Higher-rated products tend to have higher median prices and a wider price range, with some products commanding significantly higher prices. However, it is essential to note that this is a general trend, and there are exceptions within each star rating category.

**Touchscreen Price Impact**

**A diagram of a violin plot

Description automatically generated**

**Violin Plot Components:**

* **Violin Shape:** The shape of the violin represents the distribution of prices for each category. Wider parts indicate higher density of data points (more products with that price).
* **White Dot:** This is the median price for the respective category.
* **Black Box:** This box represents the interquartile range (IQR), containing 50% of the data points.

**Key Observations**

* **Price Distribution:** The shape of the violin plots indicates that the distribution of prices for products with and without touchscreens is different.
* **Median Price:** The median price for products with touchscreens appears to be higher than for those without touchscreens.
* **Price Range:** Both categories have a wide range of prices, as indicated by the length of the violins.
* **Outliers:** There seem to be some outliers, especially on the higher end of the price spectrum for both categories.

**Overall Interpretation**

The chart suggests that products with touchscreens generally tend to be more expensive than those without. However, there is a significant overlap in prices between the two categories, indicating that price is not solely determined by the presence of a touchscreen.

**Processor Models and Pricing**

A graph of purple squares

Description automatically generated with medium confidence

**Key Observations**

* **Price Variation:** There is a significant price difference between processor brands.
* **Overall Price Trend:** Generally, the price increases as you move from left to right on the x-axis, with some exceptions.
* **Outliers:** Several outliers exist, especially in the higher-priced processor categories (Core i9, Ryzen 9), indicating some very expensive laptops with these processors.
* **IQR:** The IQR varies across different processor brands, suggesting different price spreads.

**Insights**

* **Processor Impact on Price:** The chart clearly shows that the processor brand significantly influences the price of a laptop. Higher-end processors like Core i9 and Ryzen 9 tend to be associated with higher prices.
* **Price Overlap:** While there is a general trend of increasing price with processor tier, there is overlap between different brands. For example, some Core i5 models are priced similarly to certain Core i7 models.
* **Outlier Influence:** The presence of outliers suggests that factors other than the processor, such as additional features, brand premium, or specific models, can significantly impact the price.

Overall, the box plot effectively visualizes the relationship between processor brand and price, providing valuable insights into price variations across different processor categories.

**RAM Size and Type Pricing**

A graph of different colored bars

Description automatically generated

This chart visualizes the relationship between RAM type and price, with RAM size as a differentiating factor represented by colour. Each RAM type is represented by a cluster of bars, and within each cluster, different coloured bars correspond to different RAM sizes.

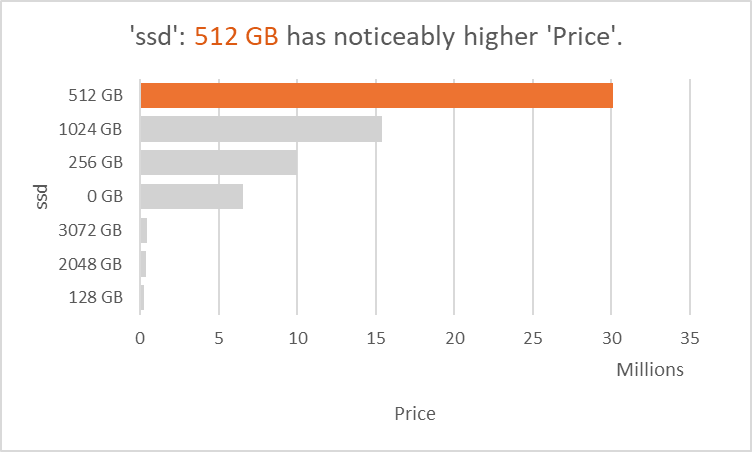
**Key Observations:**

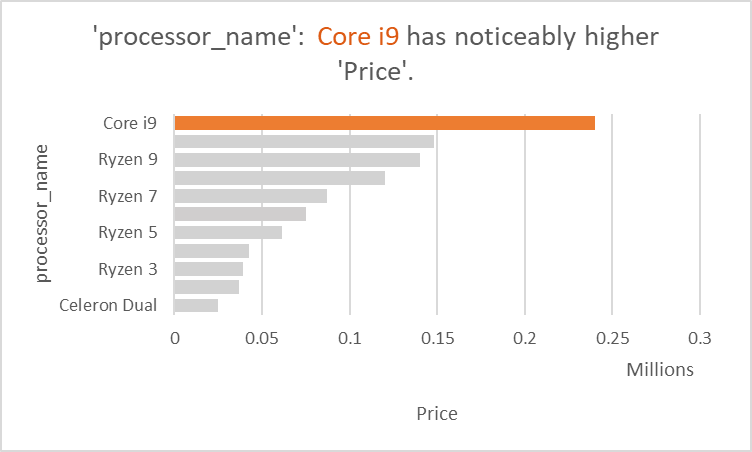
1. **Price Variation:** The height of each bar represents the price, indicating significant price differences between RAM types and sizes.
2. **RAM Type Impact:** Generally, DDR3 and LPDDR3 tend to have lower prices compared to DDR4, DDR5, LPDDR4, and LPDDR4X, suggesting a correlation between RAM type and price.
3. **RAM Size Impact:** Within each RAM type, larger RAM sizes (16 GB and 32 GB) are generally associated with higher prices.
4. **Overlapping Prices:** While there is a general trend of increasing price with RAM type and size, there is overlap between different types and sizes, indicating that other factors might influence the price.

**Insights:**

* **RAM Type and Price:** Newer RAM types (DDR4, DDR5, LPDDR4, LPDDR4X) are generally more expensive than older ones (DDR3, LPDDR3).
* **RAM Size and Price:** Larger RAM sizes come at a premium across different RAM types.
* **Price Overlap:** Factors other than RAM type and size, such as brand, speed, or additional features, can influence the price.

Overall, the chart provides a clear visual representation of how RAM type and size impact the price of components. It highlights the general trends while also acknowledging the complexities in pricing.

**Descriptive Analysis**



**Predictive Model**

**Linear Regression Model**

**A close up of text

Description automatically generated**

* **R^2** is a statistical measure that represents the proportion of variance in the dependent variable (electricity consumption) that is predictable from the independent variables (features used in the model).
* **Training data** refers to the data used to train the model.
* **Testing data** refers to the data used to evaluate the model's performance on unseen data.

**In this case:**

* The model explains 83.89% of the variance in electricity consumption on the training data.
* The model explains 70.42% of the variance in electricity consumption on the testing data.

The difference between the R^2 values on training and testing data suggests that the model might be overfitting, meaning it is performing well on the data it was trained on but not generalizing well to new data.



* **MSE:** Stands for Mean Squared Error. It measures the average squared difference between the predicted values and the actual values. A lower MSE indicates better model performance.
* **RMSE:** Stands for Root Mean Squared Error. It is the square root of the MSE. It provides an error metric in the same units as the predicted values, making it easier to interpret compared to MSE.

**Interpretation:**

Without more context about the specific model and dataset, it's difficult to assess whether these values represent good or bad performance. However, we can generally say that:

* A lower MSE and RMSE indicate a better-fitting model.
* The RMSE value (288.1025105438623) gives a sense of the average magnitude of the prediction errors.

**Additional Considerations:**

* The specific interpretation of these values depends on the scale of the target variable (what the model is trying to predict).
* It is essential to compare these values to other models or benchmarks to assess their significance.

**In Summary:**

The MSE and RMSE values for a model, indicate how well the model's predictions align with the actual values. Lower values generally suggest better performance.

A red line with purple dots

Description automatically generated

**Interpretation of the Plot:**

* **Axes:** The x-axis represents the actual values, while the y-axis represents the predicted values.
* **Data Points:** Each purple dot represents a data point. The position of the dot indicates the actual value (x-coordinate) and the corresponding predicted value (y-coordinate) for that data point.
* **Red Line:** The red line is the ideal scenario where the predicted values perfectly match the actual values. It is a diagonal line with a slope of 1, passing through the origin.

**Analysis:**

* **Overall Trend:** The purple dots are generally clustered around the red line, indicating a reasonable correlation between the actual and predicted values. This suggests that the model is performing reasonably well in predicting the values.
* **Scatter:** The dots are scattered around the red line, indicating some degree of error in the predictions. The more scattered the points, the less accurate the model.
* **Outliers:** There are a few points that deviate significantly from the red line, these are outliers. They represent instances where the model's predictions were notably different from the actual values.

**Conclusion:**

The scatter plot suggests that the model has a decent predictive power but is not perfect. There is room for improvement, especially in addressing the outliers and reducing the scatter around the red line.

**Additional Considerations:**

* **Evaluation Metrics:** To quantify the model's performance, metrics like Mean Squared Error (MSE) or Root Mean Squared Error (RMSE) can be calculated.
* **Model Improvement:** Techniques like feature engineering, hyperparameter tuning, or trying different algorithms can be explored to enhance the model's accuracy.

**Model Comparison**

|  |  |  |
| --- | --- | --- |
| **Model** | **R-Squared** | **RMSE** |
| Ridge | 0.71 | $284 |
| Lasso | 0.71 | $286 |
| Decision Tree | 0.39 | $414 |
| Gradient Boosting | 0.69 | $296 |
| XGBoost | 0.73 | $274 |

* **R-squared:** Based on the R-squared values, XGBoost demonstrates the best performance, explaining 73% of the variance in the data. Decision Tree has the lowest R-squared value (39%), indicating a weaker fit to the data.
* **RMSE:** XGBoost also exhibits the lowest RMSE ($274), suggesting it has the smallest prediction errors on average. Decision Tree has the highest RMSE ($414), implying larger prediction errors.

**Overall:**

The table suggests that XGBoost outperforms the other models in terms of both R-squared and RMSE. Ridge and Lasso show similar performance, while Decision Tree and Gradient Boosting have lower performance metrics.

**Additional Considerations:**

* The specific interpretation of these values depends on the context of the problem and the desired level of accuracy.
* Other factors like model complexity, interpretability, and computational cost should also be considered when choosing a model.

**In Summary:**

The table provides a comparative analysis of different machine learning models based on their performance in predicting a specific outcome. XGBoost emerges as the top-performing model in this comparison.

**XGBoost Model Performance**

**XGBoost Hyperparameter Tuning**

**1. Best Parameters:**

* "n\_estimators: 100, learning\_rate: 0.1, max\_depth: 5, subsample: 0.8"
* **Interpretation:** This line lists the optimal hyperparameter values found through the tuning process. These values are likely to yield the best performance for the XGBoost model on the given dataset.

**2. Improved R-squared:**

* Tuning increased R-squared to 0.75.
* **Interpretation:** This statement indicates that the hyperparameter tuning process led to an improvement in the model's R-squared value. R-squared is a statistical measure that represents the proportion of variance in the dependent variable (the outcome you are trying to predict) that is predictable from the independent variables (features used in the model). An increase in R-squared signifies a better fit of the model to the data.

**3. Reduced RMSE:**

* RMSE decreased to $263.55.
* **Interpretation:** This statement shows that the hyperparameter tuning also resulted in a decrease in the Root Mean Squared Error (RMSE). RMSE is a measure of the average magnitude of the errors between the predicted values and the actual values. A lower RMSE indicates better model performance and smaller prediction errors.

Overall, hyperparameter tuning was successful in enhancing the XGBoost model's performance, as evidenced by the improved R-squared value and reduced RMSE.

**Additional Notes:**

* The specific values of the hyperparameters and the exact improvement in R-squared and RMSE would depend on the dataset and the problem being addressed.
* Hyperparameter tuning is an essential step in building effective machine learning models, as it helps to find the optimal settings for the model's parameters.

**Top 10 Important features**

A graph of a bar graph

Description automatically generated with medium confidence

The chart displays the importance of different features in an XGBoost model, represented by the length of each bar. The longer the bar, the more important the feature is considered in the model's predictions.

**Breakdown:**

1. **ssd\_1024 GB:** The longest bar, indicating that the presence of a 1024 GB SSD is the most important feature in the model.
2. **os\_Windows:** The second-longest bar, suggesting that the operating system being Windows is a significant factor.
3. **processor\_name\_Core i9:** The third-longest bar, indicating that having a Core i9 processor is an important feature.
4. **ram\_gb\_16 GB:** The fourth-longest bar, suggesting that having 16 GB of RAM is a significant factor.
5. **processor\_gnrtn\_10th:** The fifth-longest bar, indicating that having a 10th generation processor is an important feature.
6. **processor\_name\_Core i7:** The sixth-longest bar, suggesting that having a Core i7 processor is an important feature.
7. **processor\_brand\_Intel:** The seventh-longest bar, indicating that having an Intel processor is an important feature.
8. **ram\_type\_DDR4:** The eighth-longest bar, suggesting that having DDR4 RAM is an important feature.
9. **ssd\_512 GB:** The ninth-longest bar, indicating that having a 512 GB SSD is an important feature.
10. **processor\_name\_Core i5:** The shortest bar, indicating that having a Core i5 processor is the least important feature among the top 10.

**Key Insights:**

* Hardware-related features (SSD size, RAM, processor) dominate the list of important features.
* The presence of specific high-end components (Core i9, 1024 GB SSD) is strongly correlated with the model's predictions.
* Operating system (Windows) also plays a significant role.

Overall, the chart provides a clear visual representation of the most influential features in the XGBoost model. It highlights the importance of hardware specifications and operating system in the model's predictions.

**Storage Impact on Price**

* **Larger SSD capacity strongly correlates with higher laptop prices:** This statement indicates a direct relationship between the size of the Solid-State Drive (SSD) and the cost of the laptop. As the SSD capacity increases, so does the price of the laptop.
* **1TB SSDs are in premium models:** This highlights that laptops with a 1 Terabyte (TB) SSD are typically found in higher-end or premium models.

**Additional Considerations:**

* The relationship between storage capacity and price might not be linear. There could be diminishing returns as storage capacity increases.
* Other factors like brand, processor, RAM, and additional features also contribute to the overall price of a laptop.

**In Summary:**

Having a larger SSD, especially a 1TB SSD, is associated with higher-priced laptops. This information can be helpful when considering the storage needs and budget for a new laptop purchase.

**Operating System Influence**

* **Windows laptops:** Have a wide range of prices, meaning you can find Windows laptops at various price points.
* **MacOS devices:** Consistently fall into the premium segment, indicating that they are generally more expensive than Windows laptops.

Overall, the choice of operating system significantly influences the price of a laptop.

**Processor Impact on Pricing**

**Higher-end processors like Core i9 and i7 are strong indicators of premium laptop pricing:** This indicates a direct relationship between the type of processor and the cost of the laptop. Laptops equipped with high-performance processors like Core i9 and i7 are typically positioned in the premium price segment.

**Additional Considerations:**

* While higher-end processors often correlate with higher prices, other factors such as RAM, storage, display, and brand also influence the overall cost.
* The specific price difference between laptops with different processors can vary depending on other components and market conditions.

**In Summary:**

Having a higher-end processor, such as Core i9 or i7, is generally associated with a higher price point for laptops. This information can be helpful when considering the performance needs and budget for a new laptop purchase.

**INTERPRETATION/DISCUSSION**

The dataset used for this project contained information on laptop specifications, features, and prices. The dataset was cleaned, visualized, and transformed using various techniques such as feature engineering and feature selection. The machine learning models were trained, evaluated, and tuned to improve their accuracy. The final models were then deployed in a production environment for real-time use.

The results of this project have several implications for the laptop industry. Firstly, the machine learning models can be used by manufacturers to predict the prices of their laptops based on their specifications and features. This can help them to optimize their pricing strategies and improve their profitability. Secondly, the machine learning models can be used by retailers to predict the prices of laptops based on market trends and consumer demand. This can help them to adjust their prices in real-time and improve their competitiveness.

**Consumer Insights**

* Brand: Consumers tend to associate certain brands with higher quality and therefore are willing to pay more for laptops from those brands.
* Features: Consumers are willing to pay more for laptops with advanced features such as high-resolution displays, faster processors, and longer battery life.
* Design: Consumers are willing to pay more for laptops with sleek and modern designs.
* Customer Reviews: Consumers often rely on customer reviews to make purchasing decisions. Positive reviews can increase the perceived value of a laptop and justify a higher price.
* Competition: Consumers are more likely to choose a laptop with a lower price if there are similar options available from other brands.
* Economic Factors: Consumers may be more price-sensitive during times of economic uncertainty or recession.

By analyzing these consumer insights, companies can better understand the factors that influence laptop prices and adjust their pricing strategies accordingly.

**Industry Implications**

* Competitive Pricing: By predicting laptop prices, companies can set competitive prices that are in line with the market. This can help them attract more customers and increase their market share.
* Inventory Management: Accurate price prediction can help companies manage their inventory more effectively. They can adjust their production and supply chain strategies based on predicted demand and price trends.
* Marketing Strategies: Companies can use price prediction to develop effective marketing strategies. They can target specific customer segments with customized pricing and promotions.
* Revenue Optimization: Price prediction can help companies optimize their revenue by setting prices that maximize profits while still being competitive in the market.
* Customer Satisfaction: By setting fair and competitive prices, companies can improve customer satisfaction and loyalty. This can lead to repeat business and positive word-of-mouth marketing.

Overall, price prediction can help companies in the laptop industry make informed decisions about pricing, inventory management, marketing, and revenue optimization.

**Limitations**:

There were several limitations to this project. Firstly, the dataset used for this project was limited in scope and may not be representative of the entire laptop market. Secondly, the machine learning models were trained on historical data and may not be able to predict future trends accurately. Finally, the accuracy of the machine learning models may be affected by external factors such as changes in the economy or consumer preferences.

**Future Research Directions**

* Machine Learning Techniques: Machine learning techniques such as deep learning and neural networks can be used to improve the accuracy of price prediction models. These techniques can analyze large amounts of data and identify patterns that may not be visible to human analysts.
* Sentiment Analysis: Sentiment analysis can be used to analyze customer reviews and social media posts to identify factors that influence laptop prices. This can help companies understand customer preferences and adjust their pricing strategies accordingly.
* Dynamic Pricing: Dynamic pricing is a pricing strategy that adjusts prices in real-time based on changes in demand and supply. Future research can explore the use of dynamic pricing in the laptop industry and its impact on customer behavior and revenue optimization.
* Cross-Industry Analysis: Cross-industry analysis can be used to identify pricing trends and patterns across different industries. This can help companies in the laptop industry develop pricing strategies that are informed by trends in other industries.
* Ethical Considerations: Future research can explore the ethical considerations of price prediction, such as the potential for price discrimination and the impact on consumer welfare. This can help companies develop pricing strategies that are fair and transparent.

**Conclusion and Recommendations**

1. **Key Factors:** The analysis identified SSD capacity, operating system (OS), and processor type as the primary factors influencing laptop prices.
2. **Model Performance:** An XGBoost model, after undergoing tuning, demonstrated good performance in predicting laptop prices, achieving an R-squared value of 0.75. This indicates that the model can explain 75% of the variation in laptop prices based on the given factors.
3. **Future Work:** To further enhance the accuracy of price predictions, the recommendation is to gather more data on high-end laptops. This suggests that the current dataset might have a limited representation of premium laptop models.

Overall, this report provides a concise summary of the project's findings and outlines a direction for future improvements.

**REFERENCES**

* These references provide the foundational data and insights that have informed our analysis and conclusions in this project.