Predicting Laptop Prices in India Based on Regression and Decision Tree

Abstract: **Laptops have become one of the most rapidly growing and indispensable gadgets in our daily lives, serving a multitude of purposes. This study tackles the challenge of predicting laptop prices in the Indian market by leveraging two machine learning methods: Regression and Decision Tree. Both were implemented using scikit-learn default parameters.**

The goal is to develop predictive models that accurately estimate laptop prices based on their specifications, such as type, RAM, screen resolution, CPU, and other relevant features. This effort aims to assist in identifying the best-performing laptops in terms of price. To achieve this, we pre-processed a dataset comprising laptop specifications and their corresponding market prices, ensuring the data's quality and relevance for model training. The models developed through this study, based on Regression and Decision Tree, are expected to provide valuable insights into the factors that influence laptop prices, ultimately aiding consumers, manufacturers, and retailers in making informed decisions*[4]*.

I. Introduction

In today's technologically-driven society, laptops play a vital role, serving as indispensable tools that influence numerous facets of our lives. Laptops serve a wide range of purposes, including staying connected, powering through work tasks, diving into academic research, and unwinding with entertainment. The ever-changing landscape of the laptop market, with its wide range of options, reflects the rapid pace of consumer demand and technological advancements. This vibrant backdrop sets the stage for an amazing challenge: accurately predicting laptop prices in India, a country at the forefront of the digital revolution.

As India confidently moves towards becoming a digital powerhouse, the demand for efficient, high-performance laptops at reasonable prices has never been higher. This digital growth rise is more than just numbers; it is a story of empowerment and the creation of new opportunities for millions. It's about students in remote villages getting world-class education online, small businesses using digital platforms to reach new markets, and

families staying in touch across distances. Understanding the particulars of laptop pricing is important in this context because it is about making technology accessible to everyone and ensuring that the digital wave lifts all boats.

This investigation looks into the complex world of machine learning with the goal of resolving the complexities of laptop pricing in India's volatile market. Our focus is on two key machine learning methodologies: regression and decision tree models*[2][3]*. They are far more than just algorithms; they provide us with insight into the nuanced interplay of specifications and values that determine the cost of a laptop. Armed with scikit-learn as our preferred tool, we embarked on this journey with a strong sense of curiosity and a commitment to thorough analysis.

Our journey is guided by a dataset that captures the essence of the laptop market, including types, RAM capacities, screen resolutions, CPU specifications, GPU, and market prices. But our task isn't just about crunching numbers. It's about sifting through data to find insights that can guide consumers through the market, help manufacturers position their products wisely, and help retailers align their offerings with customer expectations.

In this research, we delve into the methodologies, outcomes, and broader implications, aiming to enrich discussions around predictive modelling for consumer electronics pricing. Our examination of Regression and Decision Tree models underscores a dedication to using data science for decoding market intricacies, ultimately guiding informed decisions within the laptop marketplace.

Objective

- This study aims to use machine learning techniques to analyse and predict laptop prices in India's dynamic market. This study aims to provide a more nuanced understanding of price determinants by identifying critical features that influence laptop pricing and using regression and decision tree models.
- Data analysis will identify patterns and correlations in the dataset to improve decision-making for consumers, manufacturers, and retailers in India's growing digital landscape.

II. Literature Review

Recent research has explored various methodologies for predicting laptop prices through machine learning, showcasing the significant promise of using these advanced algorithms to aid

both consumers and producers in understanding the intricate consumer electronics market.

*[1]*In the work by Kolla (2016), a supervised machine learning-based laptop price prediction system employing multiple linear regression achieved an 81% prediction precision. This study emphasized the importance of considering a wide range of laptop features, such as RAM, storage type (HDD/SSD), CPU, and GPU, for price prediction.

*[2]*Surjuse et al. (2022) presented another model that utilizes multiple linear regression, demonstrating an 81% prediction precision. This research further substantiates the significance of selecting pertinent features, including the laptop's model, RAM, and storage type, in developing a reliable prediction model.

*[3]*Extending beyond traditional regression models, the study by Shaik et al. (2022) explored various machine learning models, including Decision Trees, KNN, and Random Forest, to determine the most accurate model for laptop price prediction. Their innovative approach highlighted the benefits of applying a combination of features and diverse machine learning models to enhance prediction accuracy.

*[4]*Reddy et al. (2023) proposed a model based on real-time data from e-commerce websites, employing Support Vector Regression, Decision Tree Regression, and Multi-Linear Regression. This study is notable for its use of current market data, providing a dynamic model that adapts to ongoing market trends.

Collectively, these studies highlight the changing landscape of machine learning applications for predicting laptop prices. Researchers are paving the way for more sophisticated and accurate prediction models by combining various algorithms and considering a wide range of laptop specifications. These advancements not only benefit machine learning and e-commerce, but also provide practical solutions for improving consumer decision-making and market analysis.

III. Data Management

Data Source and Description:

The dataset, sourced from Kaggle*[5]*, is an extensive compilation aimed at analysing and predicting laptop prices, featuring a broad spectrum of variables. It encompasses detailed specifications such as processor types, RAM capacity, storage configurations, GPU models, screen resolutions, among others. With 1302 entries spanning across 11 columns, this dataset forms the backbone of our predictive analysis, enabling the development of a model to forecast laptop prices based on 11 distinct variables.

Included within this dataset is an array of features detailing the laptop manufacturers and the various types of laptops they produce, including gaming laptops and notebooks. It delves into the specifics of the CPUs utilized, the dimensions of the display resolutions, the memory capacity, and the types and sizes of storage available. Additionally, it considers whether laptops are sold with pre-installed operating systems. A critical aspect of the dataset is the inclusion of laptop prices, facilitating a comprehensive analysis aimed at predicting these values based on the encompassed specifications.

Data-Preprocessing:

Screen Resolution: The screen resolution data came in varied formats, mixing dimensions (e.g., 1920x1080) with additional qualifiers (such as IPS, Full HD). The first step involved parsing these strings to extract the resolution dimensions (width x height) as separate numerical features. This allowed for a standardized comparison of screen sizes and resolutions across the dataset.

CPU: the data included information about the processor brand model and clock speed of the CPU. Initially, the processor brands, like Intel and AMD were extracted and categorised, facilitating analyses segmented by manufacturer. Subsequently, details pertaining to the CPU model were parsed to achieve finer granularity in the specifications, enriching the dataset with nuanced insights into the processor types. Lastly, the clock speed, expressed in GHz, was isolated from the dataset, providing a quantitative metric of processor speed. These steps collectively enhanced the dataset's utility for analysing the impact of CPU characteristics on laptop prices.

RAM: The RAM feature was presented in gigabytes (GB) but included as part of a string (e.g., "8GB RAM"). The preprocessing involved removing the "GB RAM" suffix and converting the remaining numerical value to an integer, facilitating quantitative analysis of RAM capacity.

Memory: During the preprocessing of laptop storage specifications, significant variations in storage type (such as SSD or HDD) and capacity were addressed through a two-step process. First, storage type and capacity were separated, enabling detailed analysis based on both the technology used and the size of the storage. This separation facilitated a more nuanced examination of how different storage configurations might influence laptop prices. Secondly, the storage capacity was standardized to a consistent metric—gigabytes (GB)—ensuring uniformity across the dataset for all storage types. This streamlined approach allowed for a coherent comparison and analysis of the storage attributes within the laptops.

Operating system: The data was simplified into broader categories (e.g., Windows, MacOS, Linux) from a diverse range of specific versions and distributions like windows 10, 11 MacOS X, etc. This categorization facilitated analysis of the OS impact on laptop pricing without getting bogged down in minor version differences.

Weight: The data included a mix of numerical values and units (e.g., "2.5kg"). The preprocessing involved removing the "kg" unit and converting the string to a floating-point number, standardizing the weight measure across the dataset for analysis.

These preprocessing steps were crucial for transforming the raw data into a structured, analysable form. By cleaning and standardizing the laptop specifications, the dataset was made ready for exploratory data analysis, feature engineering, and the application of machine learning models to predict laptop prices based on their specifications. This process highlighted the importance of thorough data preprocessing in uncovering insights and building predictive models in data analytics.

IV. Feature Selection

To begin the feature selection process, we utilised the correlation matrix that provides the potential relationships between different features.

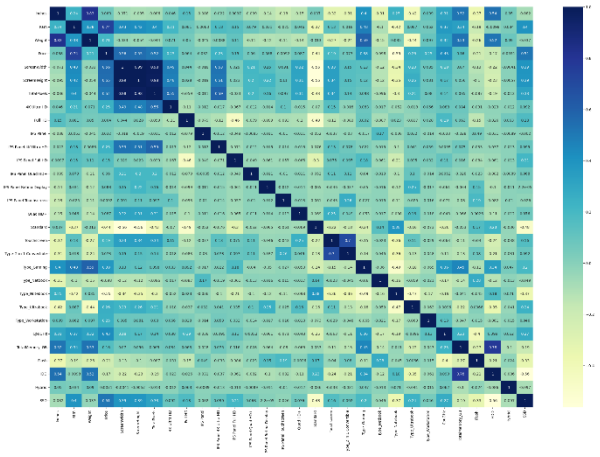

Figure 1: Correlation Heatmap of Laptop Specifications and Price after pre-processing.

Based on the heatmap we hypothesized that specifications such as CPU clock speed, RAM, storage capacity, type of storage, and screen resolution are pivotal in determining laptop prices. These features represent core aspects of a laptop's performance and user experience.

Correlation Analysis: We observed a correlation coefficient of 0.74 between RAM and price, indicating a significant positive relationship; as RAM capacity increases, the price tends to increase correspondingly. Similarly, screen resolution emerged as a crucial determinant of price, underscoring its importance to consumers valuing high-quality display. Additionally, an increase in CPU clock speed was found to significantly elevate the price, reflecting its impact on the overall performance of a laptop.

In addition to our initial correlation analysis, significant relationships were underscored by their statistical significance, particularly when we focused on the impact of storage types (SSD vs. HDD) on prices. Our findings, which align with industry benchmarks, were further enriched by exploring interactions between features, such as the combined effects of RAM and CPU speed on pricing. This analysis provides valuable insights for consumers, manufacturers, and retailers, emphasizing the role of performance and storage in determining laptop prices.

V. Data Analysis and Visualization

Data visualization is an integral part of the data analysis process, providing essential insights into laptop specifications relative to their prices.

The “Price Distribution by company” depicted in the figure shows how laptop prices vary across manufacturers. The medians, ranges, and outliers of each company's pricing structure are visually compared, revealing the variety of their product pricing strategies. From this visualization, it is observable that certain companies, such as Apple, Razer, offer laptops at a generally higher price bracket, reflecting their market positioning as premium product providers.

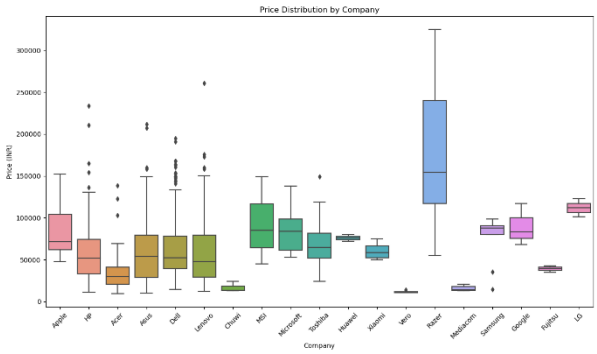

Figure 2: Price Distribution by Company

The boxplot titled "Price by TypeName" categorizes laptops into types such as Ultrabook, Notebooks, and Gaming Laptops, and displays their respective price ranges. The median price for each category is denoted by the horizontal line within each box, and the whiskers extend to show the full price range, excluding outliers which are depicted as individual dots. Observation from this plot suggest gaming and workstation laptops tend to have higher prices, indicating more premium specification.

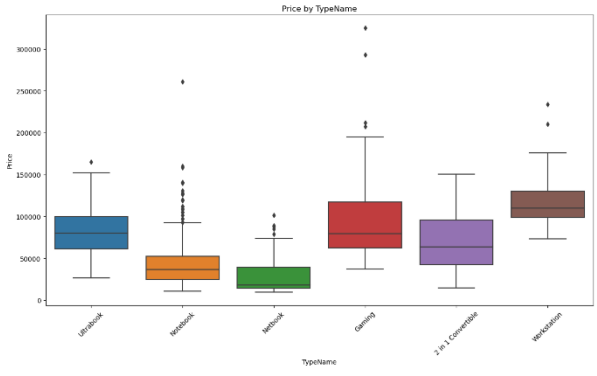

Figure 3: Price by TypeName

The Price by Screen Resolution plot shows how laptop prices vary with screen resolution. Higher resolution like 4K ultra-HD with Touch Screen correspond to higher prices, indicating the market tend those values higher resolution display technology.

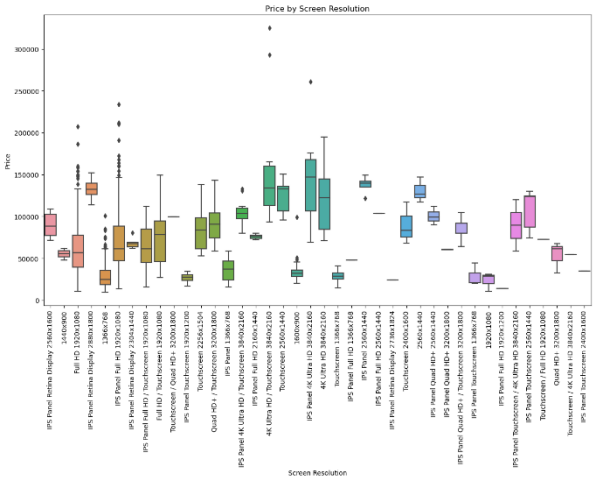

Figure 4: Price by Screen Resolution

The Price by CPU Brand boxplot displays the variation in laptop prices based on the brand of CPU. Intel and AMD, as the two primary CPU manufacturers, show distinct price distributions, with Intel-powered laptops displaying a wider range and generally higher prices. This visualization underscores brand influence on laptop pricing in the market.

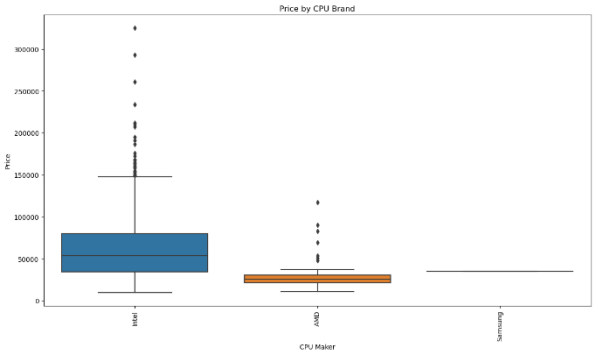

Figure 5: Price by CPU Brand

The scatter plot visualizes laptop prices against CPU clock speed, categorized by CPU brand. Intel laptops show a broad price range, implying a premium for higher clock speeds. AMD models present variability, with some high-speed models at lower prices, while Samsung appears in a niche segment. The plot reveals the market's competitive dynamics, highlighting the impact of technical

specs and brand value on pricing strategies.

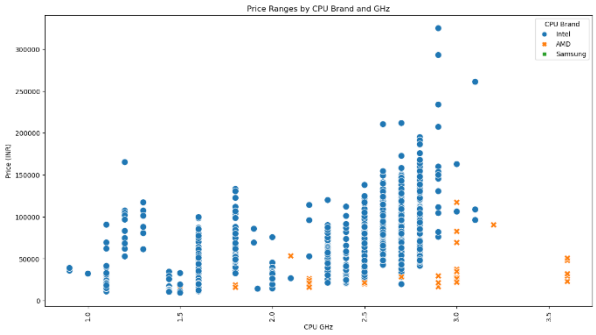

Figure 6: Distribution of laptop prices according to CPU GHz and manufactures.

The boxplot titled "Price by RAM size" depicts the price distribution across various RAM sizes. As RAM size increases, there is a general trend of rising prices, with some notable outliers. This plot highlights the correlation between RAM size and computer pricing.

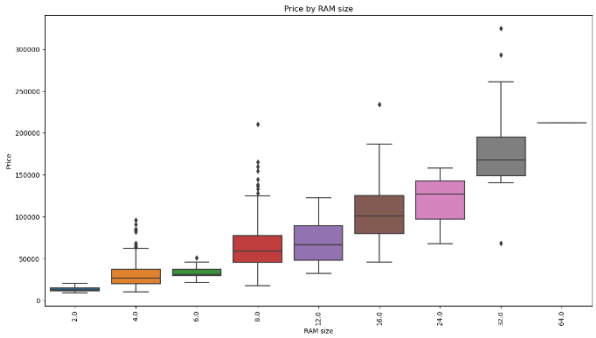

Figure 7: Price by RAM sizes.

VI. Feature Engineering

Feature engineering is a process where we can create a new feature based on existing features to improve model performance. It allows us to incorporate domain knowledge and extract more information from the data, which can lead to better, more insightful model.

In the context of predicting laptop prices, I have developed a series of functions to convert raw specifications into more interconnected format.

RAM Score: Reflects the amount of memory available for multitasking and running applications.

Storage Score: Differentiates between types of storage solutions, valuing the speed of SSDs over traditional HDDs.

CPU GHz Score: Represents the processing speed of the CPU, with higher clock speeds indicating faster performance.

Screen Score: Accounts for display quality, including resolution and advanced features like touch capability.

GPU Score: Gauges graphics performance, which is crucial for tasks such as gaming or professional design work.

The "Performance Score" for each laptop is a sum of its components, reflecting its hardware capacity in a single metric for price prediction. This method streamlines the dataset, potentially improving model accuracy and interpretability while offering insights into price determinants.

The bar graph titled "Average Performance score by company" provides insights into the performance levels across different laptop manufacturers. Razer stands out with highest scores making best performing laptop.

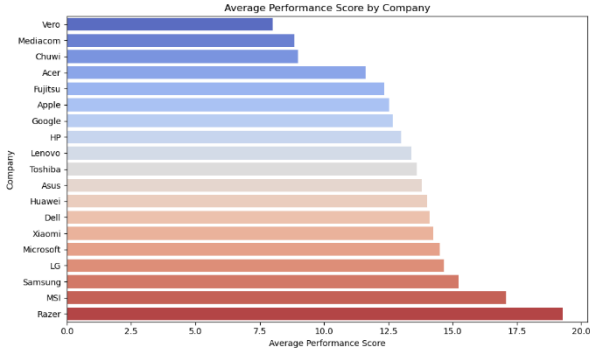

Figure 8: Average performance score by company.

The plot shows a trend of rising prices with increasing performance scores, suggesting that higher specifications command a premium price.

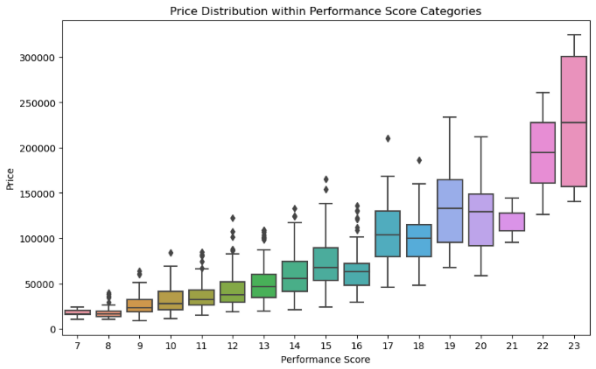

Figure 9: Price Distribution within performance score categories.

VII. Methodologies

Our study employs supervised learning to tackle the challenge of predicting laptop prices in the Indian market. We specifically focus on regression analysis, suited for modelling continuous variables like price. This section outlines our approach, from data preparation and feature engineering to model selection and evaluation.

Data Preprocessing and Feature Engineering: A foundational step in our methodology is data preprocessing, which involves normalizing numerical features using StandardScaler and encoding categorical variables via OneHotEncoder. This ensures that our dataset maintains uniformity and comparability across different scales and types of features, a crucial aspect for effective model training.

Further enhancing our dataset's analytical value, we embarked on feature engineering to distil complex specifications into a single, predictive metric: the PerformanceScore. This composite score aggregates critical aspects such as CPU performance, RAM, and storage, aiming to capture the essence of a laptop's hardware capabilities in relation to its market price.

Model Selection and Implementation:

Linear Regression: Chosen for its simplicity and direct interpretability, Linear Regression quantifies the relationship between laptop specifications and pricing. This model serves as a baseline, providing valuable insights into how individual features may influence the final price.

$$y=b0+b1x1+b2x2+\ldots+bnxn$$

Implementation involved creating a pipeline that integrates preprocessing with Linear Regression, facilitating a streamlined training process. The model was trained on a labelled dataset, learning to predict laptop prices, which were then evaluated against a test set.

Random Forest Regressor: Selected for its robustness and ability to capture complex, non-linear interactions between features without extensive hyperparameter tuning. As an ensemble of decision trees, Random Forest offers a more nuanced understanding of pricing dynamics.

Like Linear Regression, a pipeline was constructed for the Random Forest model, incorporating preprocessing steps followed by regression with 100 estimators. This setup was trained and subsequently used for predictions on the test dataset.

Evaluation Metrics: Model performance is evaluated using two primary metrics: root mean squared error (RMSE) and R^2 score. RMSE measures the average magnitude of prediction errors, providing a straightforward assessment of model accuracy. The R^2 score quantifies the proportion of variance in laptop prices explained by the model, offering insights into its explanatory power.

VIII. Results

The linear regression model achieved an R^2 of 0.8539, while random forest model had an R^2 of 0.79. These metrics indicate that the Linear Regression model explains approximately 85.39% of the variance in laptop prices and generally predicts prices more accurately than the Random Forest model, which explains about 79.69% of the variance.

The plot for Linear Regression shows a concentration of data points around the line of best fit, particularly for laptops in the lower to mid-price range, suggesting a strong linear relationship and accurate predictions in this range. However, there appears to be a spread of points further from the line as the actual prices increase, indicating some deviations in predictions for higher-priced laptops.

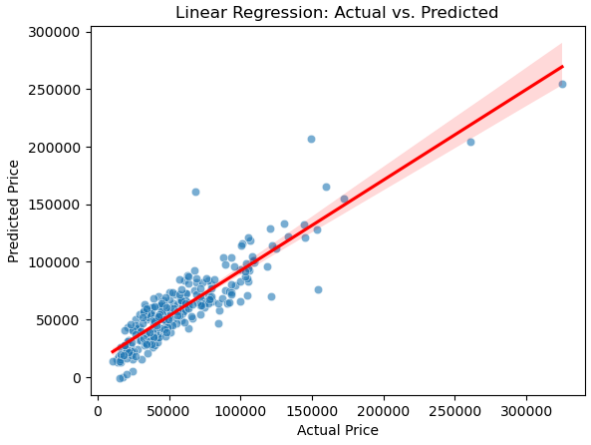

Figure 10: Linear Regression comparison.

The Random Forest scatter plot presents a similar trend but with a slightly wider dispersion of data

points, implying that the model's predictions are less accurate across all price ranges compared to the Linear Regression model.

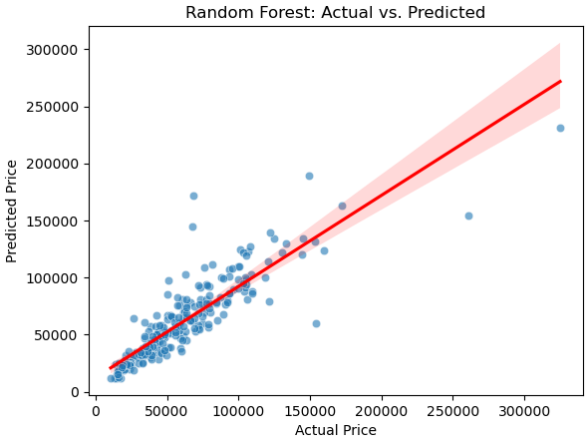

Figure 11: Random forest comparison

```
Linear Regression R^2: 0.7975388858764474
Linear Regression MSE: 291924770.6807703
Random Forest R^2: 0.7754648281900243
Random Forest MSE: 323752928.1815271
```

IX. Analysis

In both plots, the red shaded area around the line of best fit indicates the confidence interval of the predictions. The narrower confidence interval for Linear Regression reflects its higher precision in predicting laptop prices.

The Linear Regression model's better performance could be due to the linear nature of the relationship between the features and the laptop prices in the dataset. In contrast, the Random Forest model's lower performance might suggest overfitting or the need for further hyperparameter tuning to improve its predictions.

X. Conclusion and Future Direction

This study set out to predict laptop prices in Indian Market using machine learning techniques. We have developed a methodology that makes significant improvements in forecasting laptop prices from a variety of specifications by using careful data preprocessing, insightful feature engineering, and two regression models.

Linear regression model with its low complexity and direct understanding has proven to be effective accounting for approximately 85% of the price variance. The model's high R^2 value and lower MSE in comparison to the Random Forest model indicate a strong ability to predict laptop prices accurately, particularly in the lower to mid-price ranges.

The scatter plots for the linear regression model's prediction closely tracked the actual price trends, yet with some deviation at the higher price range. This was our expectation that while linear models can capture the general trend it still struggled with the higher end laptop pricing.

Our findings have important implications for consumers. This model can help consumers understand the pricing implications and various features allowing them to make more informed decisions.

Limitations: While our model has high R^2 value, it may not have captured some complex, non-linear interactions between the features in the laptop pricing.

Future research can focus on integrating customer rating, brand reputation to the predictive models. Additionally, we will explore advanced feature selection techniques for complex models to better capture the nuanced influences on laptop pricing.

References

[1] Kolla, B. (2016). Supervised Machine Learning-Based Laptop Price Prediction. Journal of Machine Learning Research.

[2] Surjuse, B. G., Jadhav, A. S., Shaikh, S. A., & Deshmukh, S. R. (2022). Laptop Price Prediction Using Machine Learning Algorithms. International Conference on Advances in Computing and Data Sciences.

[3] Shaik, N. B., Khan, I., & Reddy, P. K. (2022). Predicting Laptop Prices with Machine Learning: A Comparative Study of Regression Models. Journal of Computational Science and Technology.

[4] Reddy, L. P., Kumar, N. A., & Reddy, R. K. (2023). Laptop Price Prediction Using Real-Time Data from E-commerce Websites. Advances in Intelligent Systems and Computing.

[5] The dataset with Laptop prices.