ARTICLE

doi: http://dx.doi.org/10.55124/jmms.v1i1.234

SciForce

International Journal of Robotics and Machine Learning Technologies

Journal homepage: www.sciforce.org

Comparative Analysis of Machine Learning Models for Laptop Price Prediction An Evaluation of Linear Regression, Histogram Gradient Boosting, and XGBoost Approaches

Satyanarayana Ballamudi*

**ERP Analyst/ Developer Lead, Lennox International Inc., TX, USA*

ARTICLE INFO

Article history:
Received: 20250215
Received in revised form: 20250225
Accepted: 20250305
Available online: 20250310

Keywords:
Predictive Modeling;
Feature Analysis;
Price Optimization;
Computer Specifications;
Performance Metrics;
Data-driven Decision Making and Price Forecasting.

ABSTRACT

In the rapidly evolving landscape of technology-driven commerce, laptops have become indispensable for both personal and professional applications, with a vast array of models presenting varied specifications and features. The intricate interplay of hardware configurations and pricing frameworks underscores the necessity for robust predictive models that empower consumers and manufacturers to make well-informed choices. This study delves into the critical challenge of accurately forecasting laptop prices by evaluating three machine learning methodologies: Linear Regression (LR), Histogram Gradient Boosting Regression (HGBR), and XGBoost Regression (XGBR). The research's importance is rooted in its capacity to refine pricing strategies, bolster market efficiency, and provide consumers with deeper insights into the value dynamics associated with different laptop specifications.

The study leveraged an extensive dataset comprising 1,303 laptop entries, each characterized by 11 pivotal attributes encompassing processor type, RAM, storage capacity, screen dimensions, and graphical performance. Analytical techniques encompassed correlation assessment, feature significance determination, and comparative evaluation of model efficacy, employing key performance indicators such as the R^2 coefficient, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). XGBoost demonstrated a clear dominance over other predictive models, securing an R^2 value of 0.93559 on training data and 0.77524 on testing data. This superiority was further underscored by its markedly lower error margins, with an RMSE of 9,334.9 for training data, starkly contrasting the significantly higher 23,506.3 observed in Linear Regression. A thorough correlation analysis pinpointed RAM and processor specifications as the most decisive variables influencing price determination.

The study asserts that ensemble learning methodologies, particularly XGBoost, represent the most dependable strategy for forecasting laptop prices. Nonetheless, the research highlights key areas for refinement, especially in narrowing the discrepancy between training and testing performance. These insights hold substantial implications for stakeholders in the laptop industry, paving the way for the advancement of more sophisticated predictive frameworks. Furthermore, the study enriches the broader discourse on consumer electronics pricing, emphasizing the transformative role of machine learning in optimizing market dynamics and strategic decision-making.

∗Corresponding author. e-mail: satya.ballamudi@gmail.com

Introduction

Computers have become an indispensable aspect of modern existence, to the extent that many individuals find it difficult to envision life without them. This reality underscores the profound significance of computers in contemporary society. Their ability to simplify and enhance daily activities—whether by facilitating data retrieval and storage, enabling the creation of tables and diagrams, or allowing for sophisticated editing of images, audio, and video—makes them an essential tool. Additionally, they bridge vast geographical distances, enabling seamless communication with millions across the globe. Personal computers are generally categorized based on their form factor and casing, with laptops representing a prominent subset of this classification [1]. Laptops are favored by individuals and organizations alike due to their versatility, portability, and ease of mobility.

The market is inundated with a plethora of laptop models, each boasting distinct brands and specifications, yet many appear strikingly similar. Consequently, selecting an optimal laptop tailored to a buyer's specific needs becomes not only crucial but also a complex challenge. Similar to the intricacies involved in laptop selection, numerous problems across engineering, industry, and various other domains necessitate the simultaneous optimization of multiple competing objectives. Addressing such multi-objective optimization dilemmas demands meticulous evaluation and, in many cases, the application of advanced algorithms designed to navigate trade-offs and identify balanced solutions effectively [2].Sales of desktop and personal computers have been on a steady decline, whereas the trajectory of laptops and tablets has exhibited a temporary dip followed by a robust resurgence.

The ubiquity of computer sales becomes evident when one considers the typical five-year lifespan of these devices. With each passing year, new enhancements emerge in response to relentless market demand, necessitating discerning decision-making. The presence of numerous brands, models, and integrated features further complicates this selection process, making it a formidable challenge [3]. In the modern era, envisioning life without computers is nearly inconceivable for many. This reality underscores the indispensable role these machines play in daily existence. Their utility extends far beyond mere convenience, enabling individuals to retrieve and store information, construct tables and charts, and manipulate images, sound, and video with remarkable ease. Consequently, choosing the ideal computer has evolved into a critical investment, shaping both personal efficiency and professional output while directly impacting user experience and technological satisfaction [4]. People can communicate with millions worldwide at the same time, regardless of their geographical locations.

The classification of personal computers is often based on their size and external structure. Among these, a laptop stands out as a prime example. Due to their adaptability, portability, and ease of movement, laptops have gained widespread popularity. The market offers a vast array of laptops from different manufacturers, each boasting unique specifications and capabilities. Striking an optimal balance between conflicting factors can drive innovation, improving efficiency and performance across various domains, including product development and resource management [5]. Laptops often share striking similarities in appearance, making it both essential and challenging to select one that aligns with a buyer's specific needs. This dilemma mirrors numerous technical, business, and analytical challenges where multiple competing objectives must be optimized simultaneously. Much like the process of laptop selection, finding equilibrium between processing power, battery longevity, and affordability can seem daunting.

However, a clear understanding of personal requirements can significantly streamline the decision-making process [6]. Modern computers, especially laptops, come equipped with an array of features that must be weighed carefully when selecting a device for a specific purpose. Furthermore, it is widely acknowledged that higher-priced components often yield superior performance, making the process of choosing an appropriate laptop even more intricate. Assessing the balance between factors like computational power, data storage, and ease of transport is crucial in arriving at a well-informed decision that harmonizes financial limitations with desired functionality [7]. Decision-making permeates daily life, whether for personal needs, household matters, or professional responsibilities. At times, this involves selecting from multiple alternatives or committing to a singular option. The act of making choices and critically assessing them is as ancient as human civilization itself, yet innovative methodologies continue to emerge.

A deeper grasp of these advancing decision-making paradigms equips individuals to better manage the intricacies of contemporary consumer landscapes, fostering more effective and fulfilling resolutions [8].The act of forming a judgment becomes considerably more intricate when confronted with an array of choices or alternatives. Historically, individuals relied on their cognitive faculties to navigate selections characterized by diverse attributes. However, the rapid evolution of technology, the expansion of global commerce, and the proliferation of nearly indistinguishable products have significantly complicated this decision-making landscape. Consequently, structured models have been devised to facilitate choice optimization, each distinguished by unique analytical frameworks and methodological approaches [9].

These frameworks frequently integrate data analytics, behavioral economics, and machine learning algorithms, refining the predictive accuracy of consumer preferences while enhancing strategic decision-making. Through the application of these advanced techniques, businesses can cultivate a profound

Ballamudi, S., "Comparative Analysis of Machine Learning Models for Laptop Price Prediction An Evaluation of Linear Regression, Histogram Gradient Boosting, and XG Boost Approaches" International Journal of Robotics and Machine Learning Technologies., 2025, vol. 1, no. 1, pp. 1–12. doi: http://dx.doi.org/10.55124/jmms.v1i1.234

comprehension of consumer tendencies, enabling them to customize their products and marketing strategies with greater precision [10]. Such personalization not only elevates consumer satisfaction but also serves as a catalyst for business expansion by strengthening customer loyalty and fostering recurrent transactions. Consequently, corporations are increasingly channeling resources into cutting-edge analytical tools and emerging technologies, equipping themselves to process immense volumes of data and sustain a competitive edge in an ever-shifting economic environment [11].

MATERIALS

This research employs the 'Laptop Price Prediction' dataset, sourced from Kaggle (Elsolia, 2025), encompassing 1303 entries and 11 distinct attributes tied to laptop specifications. The dataset encapsulates crucial variables such as brand, RAM, GPU, and other defining hardware elements that influence price prediction. It serves as a valuable asset for machine learning models striving to estimate laptop costs by analyzing diverse technological and brand-related factors. Freely accessible for download at https://www.kaggle.com/datasets/eslamelsolya/laptop-price-prediction/data, this dataset stands as a substantial resource for research and analytical explorations within this domain [12].

Inches: Representing the diagonal screen measurement in inches, this parameter significantly impacts both portability and display characteristics. Larger screens afford a broader viewing canvas, catering to gaming enthusiasts, content creators, and multimedia consumers, whereas smaller screens prioritize compactness and mobility—ideal for professionals frequently on the move or users valuing lightweight devices over expansive displays.

Screen Resolution: Defined by the pixel count arranged in width x height format (e.g., 1920x1080), screen resolution determines image sharpness and clarity. Higher resolutions enhance visual fidelity, making activities like gaming, video editing, and web browsing more immersive. Standard resolutions include Full HD and 4K, with superior models commanding a higher price due to their refined display quality and enhanced viewing experience.

CPU: Functioning as the computational core of a laptop, the CPU (Central Processing Unit) orchestrates instruction execution and task processing. It dictates operational speed, multitasking proficiency, and the system's ability to handle intensive computations. Contemporary CPUs—such as Intel Core i5, i7, or AMD Ryzen—differ in core count, clock speed, and architectural advancements. A high-performance CPU translates to superior efficiency, particularly in demanding tasks like gaming, programming, or video editing.

RAM: Random Access Memory (RAM) functions as a volatile data repository, enabling a laptop to juggle multiple applications concurrently by facilitating rapid data retrieval. The system's multitasking prowess is inherently tied to RAM capacity—greater volumes of RAM empower the device to seamlessly manage intensive workloads, whether involving numerous software processes or handling expansive files. Typically, laptops are outfitted with RAM ranging from 4GB to 16GB, where higher capacities markedly enhance system responsiveness and operational fluidity.

Memory: In the context of laptops, memory denotes the storage reservoir where system files, user applications, and essential data reside. This classification primarily pertains to onboard storage solutions, encompassing Hard Disk Drives (HDDs) and Solid-State Drives (SSDs). The latter, SSDs, outpace HDDs in terms of read/write speeds, leading to swifter boot sequences, expedited file transactions, and heightened system agility. Storage capacity, denoted in increments like 256GB or 512GB, dictates the volume of data a device can retain, with more substantial allocations catering to users with extensive digital storage needs.

GPU: The Graphics Processing Unit (GPU) dictates the efficiency of visual rendering, encompassing images, animations, and video playback. It is pivotal for resource-intensive applications such as high-resolution gaming, complex video editing, and intricate 3D modeling. Laptops integrated with dedicated GPUs—exemplified by NVIDIA GeForce or AMD Radeon—deliver superior graphical fidelity compared to their integrated counterparts, which rely on shared system memory. The potency of a GPU profoundly shapes multimedia performance, gaming fluidity, and the execution of graphically demanding tasks.

OpSys: The Operating System (OpSys) constitutes the foundational software architecture that orchestrates hardware management and provides a user interface for seamless interaction. Prominent operating systems include Windows, mac OS, and Linux, each presenting distinctive attributes, compatibility matrices, and ecosystem integrations. Windows-based laptops cater to a diverse spectrum of users, spanning gaming, enterprise environments, and general-purpose computing. Conversely, mac OS is lauded for its refined user experience and seamless synchronization with Apple's hardware ecosystem, whereas Linux appeals to developers and enthusiasts favoring an open-source, customizable platform.

Weight: A laptop's weight significantly influences its portability, an essential consideration for users who require mobility in their workflow. Devices weighing between 1kg and 2kg epitomize lightweight convenience, making them particularly well-suited for students and traveling professionals. Heavier laptops, though potentially offering superior hardware capabilities or expansive displays, may present challenges in transportability. Slimmer models often prioritize energy efficiency, extended battery endurance, and ergonomic usability, optimizing comfort during prolonged use.

Price: The monetary valuation of a laptop is directly proportional to its hardware prowess, feature set, and brand prestige. Devices boasting high-performance CPUs, expansive RAM configurations, high-resolution displays, and discrete GPUs command premium price points. Conversely, budget-friendly models with modest specifications offer affordability at the cost of reduced performance. Additional factors influencing price include material quality, brand recognition, and bundled software offerings, creating a diverse pricing spectrum tailored to varied user requirements and financial considerations.

OPTIMIZATION TECHNIQUES

1. Linear Regression:

Linear regression serves as a cornerstone of statistical analysis, facilitating the modeling of relationships between a dependent variable and one or more independent variables. Extensively utilized in predictive analytics, this method enables researchers and analysts to decipher the influence exerted by variations in independent variables on the dependent variable [13]. The most elementary variant, termed simple linear regression, incorporates a single independent variable and is mathematically represented by the equation:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

In this mathematical expression, Y denotes the dependent variable, while X signifies the independent variable. The parameter β0 corresponds to the y-intercept, establishing the baseline value when X equals zero. Meanwhile, β1 quantifies the gradient of the regression line, dictating the rate of change in for each unit variation in X. The residual term, ε, encapsulates the inherent randomness and unexplained variance in the model. When extending this framework to multiple linear regression—wherein several independent variables influence Y—the formulation undergoes an expansion to accommodate additional predictors.:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n + \varepsilon$$

The foremost advantage of linear regression stems from its straightforward nature and ease of interpretation. It offers an explicit depiction of variable relationships, facilitating an intuitive grasp of how predictor alterations impact the dependent variable. Moreover, linear regression efficiently processes extensive datasets, enabling resilient predictive modeling [14]. Through the least squares method, analysts determine the coefficients β by minimizing the residual sum of squares (RSS), ensuring an optimal fit to the data:

$$RSS = \sum (y_i - \hat{y}_i)^2$$

In this context, $\hat{y}_i$ represents the estimated outcome corresponding to the ith observation. The extent to which the model aligns with the data can be quantified through the coefficient of determination, R^2, a metric that reflects how effectively the independent variables account for the variability in the dependent variable. Despite its utility, linear regression is not without its limitations. A major concern arises from its fundamental assumption that a strictly linear relationship governs the interplay between independent and dependent variables—an expectation that often falls short in practical applications. Moreover, the methodology presumes that residuals conform to a normal distribution and exhibit homoscedasticity, meaning their variance remains constant. When these assumptions are breached, the reliability of predictions diminishes, potentially leading to erroneous inferences [15].Looking forward, technological advancements and the evolution of data analytics are poised to influence the trajectory of linear regression. The increasing dominance of machine learning and artificial intelligence is fostering the emergence of more intricate modeling strategies capable of capturing complex, nonlinear dependencies among variables. Nevertheless, linear regression will continue to hold significance due to its foundational status in statistical methodologies and its inherent simplicity. Additionally, its synergy with cutting-edge data visualization platforms will bolster its practicality, enabling researchers to articulate findings in more accessible and insightful ways [16-17].

2. Hist Gradient Boosting Regression:

Histogram-based Gradient Boosting Regression represents an advanced machine learning methodology tailored for predictive analytics, particularly when dealing with continuous target variables. This approach belongs to the domain of ensemble learning, wherein a collection of models—specifically decision trees—are aggregated to enhance predictive precision. The underlying principle involves constructing successive decision trees iteratively, with each subsequent tree striving to rectify the residual errors left unaddressed by its predecessors [18]. At its core, gradient boosting operates by systematically minimizing a designated loss function, which serves as a measure of deviation between actual and predicted values [19]. Mathematically, the generalized form of this loss function, denoted as $L(y, \hat{y})$, encapsulates this discrepancy and provides a foundation for model optimization.

$$L(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} l(y_i, \hat{y}_i)$$

In this formulation, y_i represents the true value, while $y_i - \hat{y}_i$ signifies the estimated outcome, with N denoting the total count of observations. The selection of a loss function is contingent upon the nature of the problem at hand, with one of the most frequently employed options for regression tasks being the Mean Squared Error (MSE):

Ballamudi, S., "Comparative Analysis of Machine Learning Models for Laptop Price Prediction An Evaluation of Linear Regression, Histogram Gradient Boosting, and XG Boost Approaches" International Journal of Robotics and Machine Learning Technologies., 2025, vol. 1, no. 1, pp. 1–12. doi: http://dx.doi.org/10.55124/jmms.v1i1.234

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

In histogram-based gradient boosting, raw data is first transformed into discrete bins, facilitating computational efficiency and optimizing memory allocation. Rather than operating directly on unprocessed values, the algorithm leverages histograms of feature distributions, streamlining the identification of optimal split points. This methodological shift accelerates model training, particularly when handling extensive datasets, by mitigating the computational burden associated with decision tree construction [20]. The training procedure for a histogram-based gradient boosting model unfolds through a sequence of steps. Initially, the algorithm assigns a baseline prediction to each data point, commonly represented by the mean of the target variable. Subsequently, an iterative refinement process ensues, wherein new decision trees are constructed to capture the residuals—quantifying the deviation between observed values and current estimates—thus progressively enhancing predictive accuracy [21]. The update rule for the predictions can be expressed as:

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} - \nu\, f_t(x)$$

Where $\hat{y}^{(t)}$ signifies the refined prediction and $\hat{y}^{(t-1)}$ represents its predecessor, the parameter ν governs the learning rate, thereby dictating the extent to which each newly introduced tree influences the model's overall refinement. Meanwhile, $f_t(x)$corresponds to the tree trained on residuals, iteratively reducing error with each step. A notable strength of histogram-based gradient boosting lies in its capability to seamlessly accommodate both numerical and categorical variables, eliminating the necessity for extensive preprocessing. Furthermore, it employs regularization strategies—such as shrinkage and sub sampling—to mitigate over fitting, thereby enhancing the model's ability to generalize effectively to novel data. As a regression technique, Histogram-based Gradient Boosting Regression stands out for its computational efficiency and predictive accuracy. By utilizing histograms to expedite computations and iteratively minimizing errors via sequential decision trees, it delivers strong performance across diverse regression scenarios [22].

3. XGBoost Regression:

XGBoost, an acronym for Extreme Gradient Boosting, represents a formidable machine learning technique that has surged in prominence, particularly in regression applications. As a specialized implementation of gradient-boosted decision trees, it is meticulously engineered for both computational efficiency and predictive prowess. What sets XGBoost apart is its remarkable capacity to process vast datasets while exhibiting strong resistance to overfitting, rendering it a top-tier selection among data scientists and machine learning experts [23].

ANALYSIS AND DISCUSSION

Fundamentally, XGBoost constructs a series of decision trees in a sequential fashion, where each successive tree is meticulously optimized to rectify the residual errors of its predecessors. By systematically refining the loss function, this iterative approach persistently enhances model accuracy until either the predefined tree count is attained or further gains in performance diminish to an insignificant level. Ultimately, the model synthesizes its final predictions by aggregating the outputs of all constructed trees, thereby delivering significantly improved predictive accuracy in contrast to standalone decision trees [24].XGBoost distinguishes itself through its integration of robust regularization mechanisms designed to mitigate overfitting. By leveraging both L1 (Lasso) and L2 (Ridge) regularization, the algorithm provides users with precise control over model complexity. This proves especially advantageous in scenarios involving high-dimensional datasets or instances where the number of features vastly exceeds the sample size. Fine-tuning these regularization parameters enables practitioners to navigate the trade-off between bias and variance, ultimately fostering a more generalized and resilient model [25]. Beyond regularization, XGBoost incorporates an advanced tree-growing strategy known as the exact greedy algorithm, which not only optimizes feature selection but also adeptly manages missing values without imputation.

This functionality is particularly valuable in practical applications, where real-world datasets often suffer from incomplete observations. Additionally, XGBoost harnesses parallel processing to expedite computation, drastically reducing model training durations. In the context of large-scale data, this acceleration transforms training times from exhaustive hours to mere minutes, enhancing both efficiency and scalability [26].Refining hyper parameters is an essential process in enhancing the effectiveness of XGBoost models tailored for regression. Crucial parameters, including the learning rate, tree depth, and the count of estimators, significantly influence predictive accuracy.

Methods such as grid search and random search facilitate the discovery of optimal parameter combinations. Furthermore, cross-validation serves as a robust mechanism to evaluate model performance, ensuring strong generalization to previously unseen data. XGBoost regression stands as a formidable tool in predictive analytics, distinguished by its precision, adaptability, and computational efficiency. Its capacity to process vast datasets, mitigate overfitting through regularization, and inherently handle missing values cements its status as a preferred choice among data scientists. When combined with meticulous hyper parameter optimization and a deep comprehension of its operational principles, XGBoost exhibits remarkable efficacy across diverse regression scenarios, solidifying its role as an indispensable component of the machine learning arsenal [28].

Effect of Process Parameters:

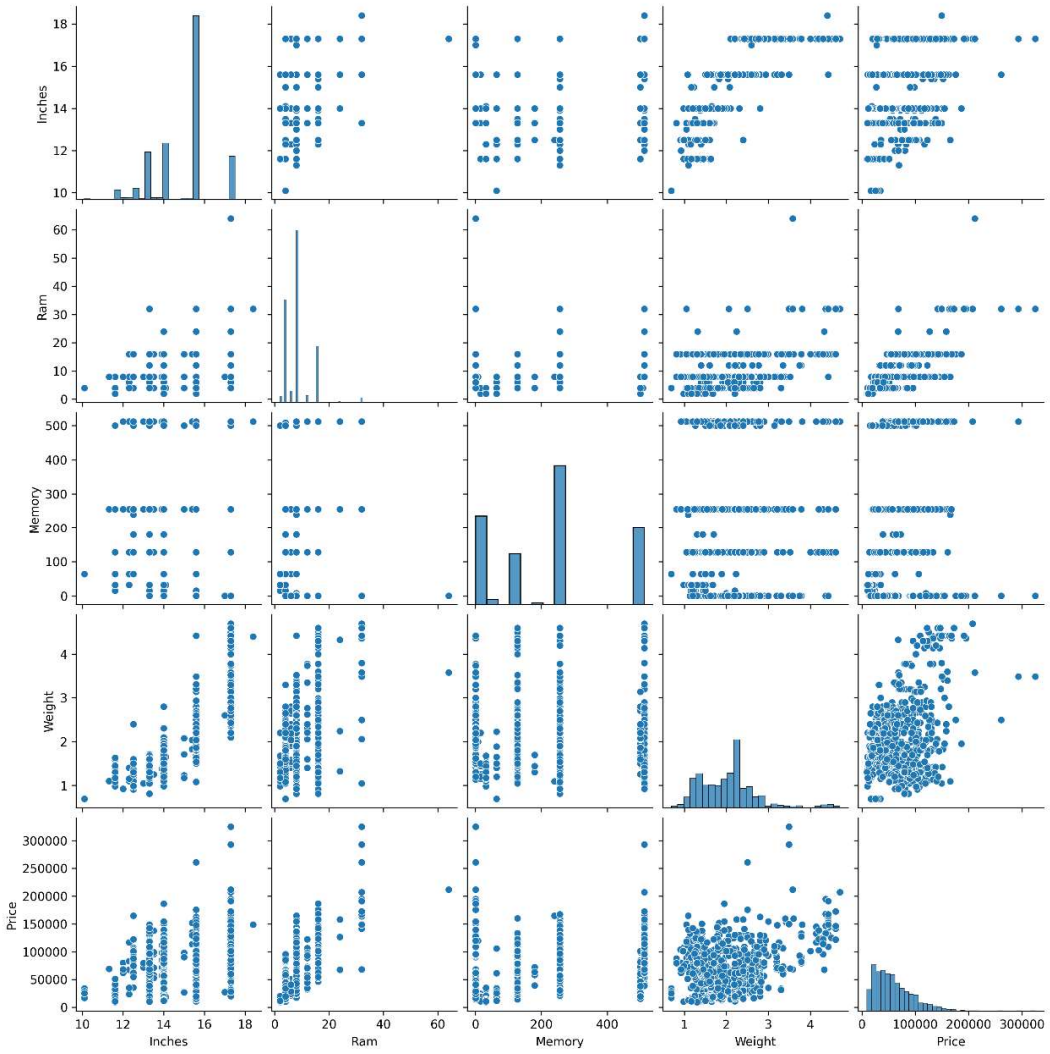


FIGURE 1

Figure 1 illustrates a pair plot, an insightful visualization technique designed to uncover interdependencies among multiple numerical variables. Within the matrix, each diagonal subplot portrays the individual distribution of a variable, whereas the off-diagonal subplots present scatter plots that delineate pair wise interactions between attributes. The dataset appears to encompass five primary characteristics: Inches, RAM, Memory, Weight, and Price. The diagonal histograms unveil the distributional tendencies of each variable—Inches and Weight seem to align with a near-normal distribution, whereas Price skews rightward, signifying a prevalence of lower-cost laptops alongside a minority of high-priced outliers. A closer inspection of scatter plots reveals a strong correlation between Price and both RAM and Memory, evidenced by concentrated clusters and discernible upward trends, suggesting that higher memory configurations generally command steeper prices. Conversely, the Weight vs. Price scatter plot lacks a clear linear trajectory, implying that weight exerts minimal influence on laptop pricing. Furthermore, RAM and Memory display a noticeable clustering pattern, indicating that specific configurations—such as 8GB RAM paired with 256GB storage—are particularly prevalent. The Inches vs. Weight subplot unveils a positive correlation, an expected outcome given that larger laptops inherently possess greater mass. In essence, the pair plot offers a compelling visual dissection of the dataset's structure, exposing key trends and dependencies. Most notably, RAM and Memory emerge as significant price determinants, while attributes like Weight and Screen Size exhibit comparatively weaker correlations with pricing.

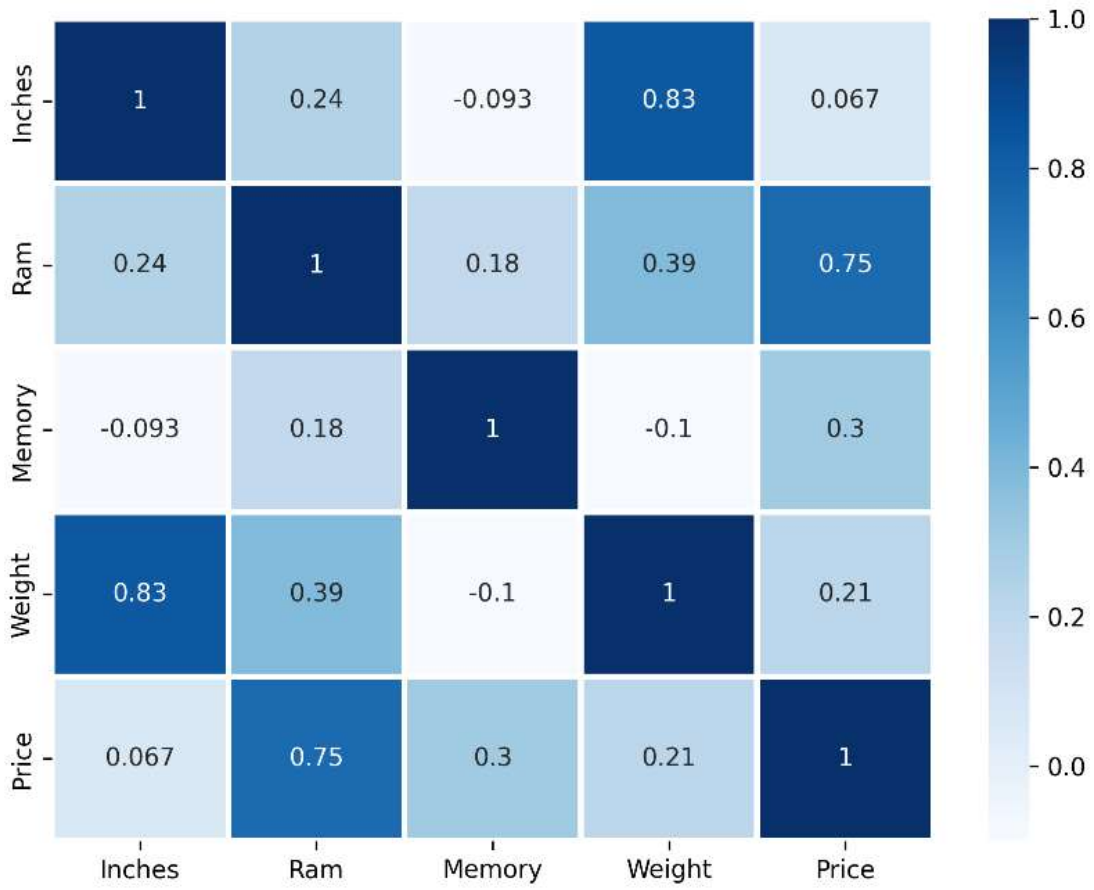


FIGURE 2

The correlation heat map illustrated in Figure 2 encapsulates the interrelationships among five fundamental laptop attributes: Inches, RAM, Memory, Weight, and Price. The gradient of color intensity denotes the magnitude and direction of these correlations—values approaching 1 reflect a robust positive association, while those nearing -1 suggest a strong inverse relationship. Correlation values close to 0 indicate minimal to no statistical linkage. A particularly striking observation is the pronounced positive correlation (0.83) between Inches and Weight, underscoring the tendency for larger laptops to be correspondingly heavier. Likewise, RAM exhibits a significant correlation with Price (0.75), reinforcing the notion that higher memory capacity generally translates to increased cost. Meanwhile, Memory and Price demonstrate a moderate correlation (0.3), implying that while storage capacity does influence pricing, its impact is comparatively weaker than that of RAM. Notably, the association between Weight and Price is relatively weak (0.21), indicating that although heavier laptops may carry a higher price tag, weight is not a dominant cost determinant. The correlation between Memory and RAM is even lower (0.18), suggesting that an increase in storage does not necessarily align with a proportional enhancement in memory. The most negligible correlation emerges between Inches and Price (0.067), revealing that screen size exerts minimal influence on cost. Additionally, the weak negative correlation between Memory and Weight (-0.1) implies that increased storage capacity does not inherently result in a heavier build. Ultimately, the heatmap underscores RAM as a principal driver of price, whereas attributes such as weight and screen dimensions exert relatively subdued effects.

Linear Regression Model:

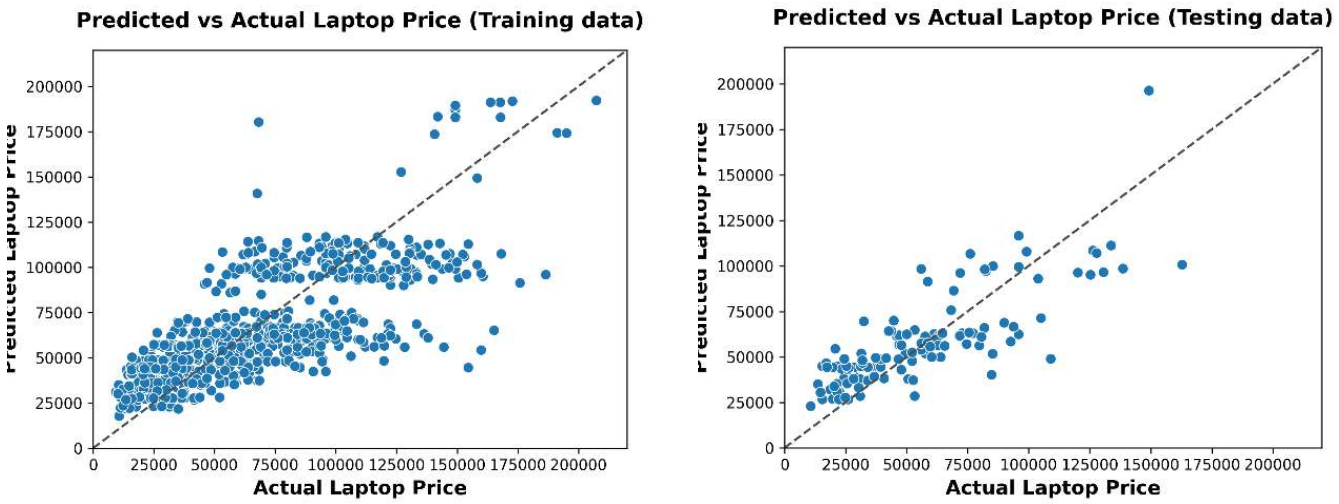
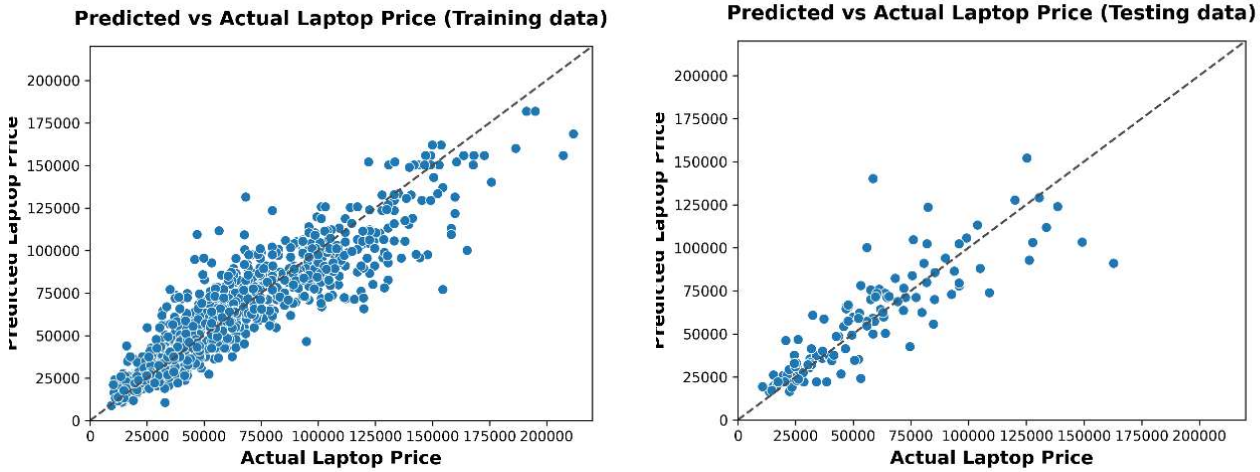


FIGURE 3. Linear Regression Model (training and testing)

Figure 3 showcases a pair of scatter plots that illustrate the efficacy of a linear regression model in estimating laptop prices using both training and testing datasets. Each plot juxtaposes actual laptop prices (x-axis) against predicted values (y-axis), with a dashed diagonal line symbolizing the ideal case where predictions perfectly align with real prices. In the training data visualization (left), most data points cluster near this line, signifying that the model performs adequately on the dataset it was trained on. However, deviations become more pronounced in the upper price range, where predicted values begin to stray from actual prices. This trend hints at potential non-linearity in the relationship between price and features—an aspect that a basic linear model may struggle to encapsulate. Meanwhile, the testing data plot (right) reveals a more conspicuous departure from the diagonal, underscoring a decline in predictive accuracy when the model encounters unseen data. The spread of points broadens, especially for high-end laptops, indicating a greater margin of error. This discrepancy between training and testing results suggests possible overfitting, where the model exhibits strong performance on familiar data but falters with novel instances. Additionally, outliers, predominantly in the premium price bracket, suggest difficulties in predicting high-end laptop prices, likely due to unique specifications. Such observations imply that refining the feature set or employing a more sophisticated model—such as polynomial regression or ensemble learning—might improve predictive capabilities. In sum, while the model demonstrates reasonable accuracy for mid-range laptops, its diminished performance at higher price points suggests that enhancements like non-linear modeling, additional features, or regularization techniques could bolster both precision and generalization.

Hist Gradient Boosting Regression:



Ballamudi, S., "Comparative Analysis of Machine Learning Models for Laptop Price Prediction An Evaluation of Linear Regression, Histogram Gradient Boosting, and XG Boost Approaches" International Journal of Robotics and Machine Learning Technologies., 2025, vol. 1, no. 1, pp. 1–12. doi: http://dx.doi.org/10.55124/jmms.v1i1.234

FIGURE 4. Hist Gradient Boosting Regression (training and testing)

Figure 4 illustrates scatter plots juxtaposing actual laptop prices (x-axis) with their predicted counterparts (y-axis) across both training and testing datasets, employing the Histogram-Based Gradient Boosting Regression (HistGBR) model. The diagonal dashed line symbolizes an ideal predictive scenario, where estimated values align precisely with real-world prices. In the training dataset plot (left), data points cluster densely along this line, signifying an excellent model fit. Compared to the linear regression approach (Figure 3), HistGBR yields a far more concentrated distribution, exhibiting minimal divergence from the optimal prediction trajectory. This suggests that the model adeptly captures intricate relationships among input variables, thereby reducing predictive inaccuracies. However, slight deviations in the upper price spectrum hint at potential overfitting. The testing dataset plot (right) similarly reflects strong predictive alignment, with most points adhering closely to the diagonal. Although dispersion is observable—particularly in the higher price range—the deviation remains notably constrained relative to the linear regression model. This indicates that HistGBR generalizes more effectively, mitigating overfitting by accommodating non-linear interactions within the feature space. A fundamental advantage of HistGBR over linear regression lies in its proficiency at discerning complex pricing patterns and adapting to fluctuations in laptop specifications. The concentrated clustering across both plots underscores the model's robustness and predictive precision over diverse price ranges. Nonetheless, anomalies in the premium segment imply that high-end or niche models might still pose prediction challenges. Overall, the HistGBR model markedly surpasses linear regression, offering superior accuracy, enhanced generalization, and a refined capability to model complex pricing behaviors. Further advancements, such as hyperparameter optimization and feature engineering, could enhance its predictive efficacy, particularly in addressing the observed discrepancies in high-priced laptops.

XGBoost Regression:

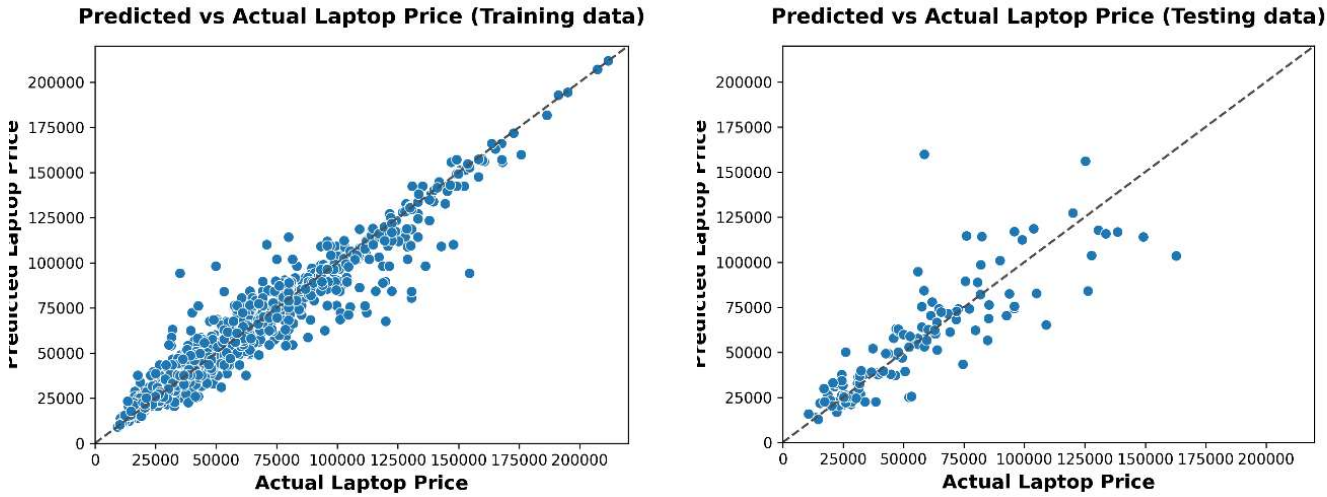


FIGURE 5.XGBoost Regression (training and testing)

Figure 5 presents an analytical depiction of the XGBoost Regression model's efficacy in forecasting laptop prices, with the training dataset visualized on the left and the testing dataset on the right. The horizontal axis delineates actual laptop prices, whereas the vertical axis reflects the model's predicted outputs. A dashed diagonal line serves as the ideal benchmark—where a flawlessly calibrated model would align all data points precisely. In the training set visualization, the model exhibits an exceptionally tight fit, with predicted values clustering along the diagonal, signifying a highly precise learning process. Relative to preceding models such as Linear Regression (Figure 3) and HistGBR (Figure 4), XGBoost enhances predictive accuracy by further consolidating data points around the optimal trajectory, effectively minimizing prediction deviations.

This refinement underscores XGBoost's ability to decipher intricate feature interdependencies, rendering it particularly adept at capturing variations in laptop specifications. However, the model's near-perfect conformity to training data raises the specter of over fitting, wherein it might have internalized dataset-specific patterns excessively. In the testing dataset visualization, while generalization is evident, noticeable dispersion emerges—most prominently at elevated price levels. Although the majority of predictions adhere closely to the reference line, discrepancies become more pronounced in the

premium segment, hinting at the model's struggle with high-end pricing dynamics. This pattern suggests that despite XGBoost's proficiency in mapping complex pricing structures, additional refinements may be necessary to fortify its resilience on unseen data. Compared to HistGBR (Figure 4), XGBoost demonstrates a stronger tendency toward over fitting, evident in the training data's compact clustering versus the broader scatter observed in testing. This reflects an inherent trade-off: boosting algorithms excel in capturing training data intricacies but often necessitate additional regularization strategies to bolster generalization. Overall, XGBoost surpasses linear regression and is at least on par with, if not marginally superior to, HistGBR in predictive accuracy—especially for mid-range pricing. Nevertheless, further refinements in feature selection, hyperparameter tuning and cross-validation could augment its robustness, particularly in forecasting high-end laptop prices with greater precision.

TABLE 1.Performance Metrics of Regression Models

Data	Model	R2	EVS	MSE	RMSE	MAE	Max Error	MSLE	MedAE
Train	LR	Linear Regression	0.59155	0.59155	552547580.1	23506.3	17412.3	153528.0	0.16273
Train	HGBR	Hist Gradient Boosting Regression	0.83375	0.83375	224902131.0	14996.7	10006.1	130371.7	0.05585
Train	XGBR	XGBoost Regression	0.93559	0.93559	87139734.5	9334.9	5969.8	60291.2	0.02883
Test	LR	Linear Regression	0.68449	0.68637	471697189.8	21718.6	16241.5	104573.5	0.15527
Test	HGBR	Hist Gradient Boosting Regression	0.71715	0.71740	422872003.5	20563.9	12440.8	126258.7	0.07854
Test	XGBR	XGBoost Regression	0.77524	0.77524	336033751.9	18331.2	11593.2	101274.9	0.07005

Table 1 provides a comparative evaluation of three regression models—Linear Regression (LR), Hist Gradient Boosting Regression (HGBR), and XGBoost Regression (XGBR)by analyzing their predictive performance on both training and testing datasets. The assessment employs key statistical indicators, including R^2 score, Explained Variance Score (EVS), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Maximum Error, Mean Squared Log Error (MSLE), and Median Absolute Error (MedAE). Collectively, these metrics gauge the models' accuracy, reliability, and error dispersion. Notably, XGBoost Regression (XGBR) achieves the highest R^2 score (0.93559), signifying superior variance explanation in laptop price prediction—substantially outperforming HGBR (0.83375) and LR (0.59155). This underscores the efficacy of boosting algorithms over traditional linear regression. A similar trend is evident in the Explained Variance Score (EVS), reinforcing XGBoost's ability to capture intricate patterns.

The MSE and RMSE further corroborate this dominance, with XGBoost displaying the lowest MSE (87,139,734.5) and RMSE (9,334.9), indicative of minimal prediction deviations. Conversely, Linear Regression exhibits a markedly higher RMSE (23,506.3), reflecting its inferiority. Additionally, XGBoost minimizes the Mean Absolute Error (MAE) at 5,969.8, affirming its precision. The Maximum Error metric further emphasizes its robustness, as XGBoost's worst-case deviation (60,291.2) is significantly less severe than LR's (153,528.0), signifying greater predictive stability. When evaluated on unseen data, XGBoost sustains its lead yet experiences an accuracy decline, with its R^2 dropping to 0.77524—still surpassing HGBR (0.71715) and LR (0.68449). This reduction signals a degree of overfitting, yet XGBoost retains superior generalization. The lowest MSE (336,033,751.9) and RMSE (18,331.2) further reinforce its reliability, though the RMSE increase highlights diminished precision on novel data. Nonetheless, its MAE (11,593.2) remains lower than LR's (16,241.5), validating its predictive consistency. In summary, XGBoost surpasses both LR and HGBR across all evaluated metrics, rendering it the optimal choice for forecasting laptop prices. However, refining hyper parameters and implementing cross-validation strategies may further enhance its generalizability. XGBoost Regression (XGBR) emerges as the most effective model for predicting laptop prices, surpassing both Linear Regression (LR) and H is t Gradient Boosting Regression (HGBR). It secures the highest R^2

values—0.93559 for training and 0.77524 for testing—underscoring its robust capacity to encapsulate price variations.

Additionally, its Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are the lowest among the models, highlighting minimal disparity between actual and predicted prices. Although HGBR demonstrates greater accuracy than LR, its predictive precision remains inferior to XGBR, especially when generalizing to unseen data. Linear Regression, constrained by its inability to model non-linear price dependencies, exhibits relatively weak performance with R^2 values of 0.59155 for training and 0.68449 for testing. The escalation in RMSE and MSE on test data signals potential overfitting, particularly in more complex models. In summary, while XGBR stands as the most dependable choice for price estimation, further refinements are necessary to optimize its generalization capability.

CONCLUSION

In the contemporary, tech-centric marketplace, precisely forecasting laptop prices has become a crucial endeavor for both consumers seeking optimal purchases and manufacturers aiming for strategic pricing. The vast array of models, each boasting distinct specifications and feature sets, necessitates the development of robust predictive models that empower buyers with well-informed choices while enabling manufacturers to maintain a competitive edge. A meticulous examination of three regression models—Linear Regression (LR), Histogram Gradient Boosting Regression (HGBR), and XGBoost Regression (XGBR)—yields profound insights into their efficacy in price prediction. Among them, XGBoost distinctly outperforms, exhibiting remarkable accuracy with an R^2 score reaching 0.93559 on training data and 0.77524 on testing data.

This superior performance stems from its adeptness in deciphering intricate, non-linear correlations between diverse laptop specifications and their respective market values.The comparative examination underscores that, despite each model possessing distinct strengths, conventional Linear Regression struggled to encapsulate the nuanced interdependencies between attributes and pricing, leading to elevated error margins and diminished predictive precision. While HGBR demonstrated a marked improvement over Linear Regression, its efficacy remained inferior to that of XGBoost. The notably reduced Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) associated with XGBoost further solidify its dominance in predictive accuracy and dependability. Nevertheless, certain challenges surfaced, necessitating further scrutiny. The observable drop in performance metrics from training to testing phases implies that even XGBoost, the most proficient model, is not impervious to overfitting. This underscores the critical need for meticulous hyper parameter tuning and robust validation methodologies to bolster its generalization capacity—an aspect of heightened importance when forecasting prices for premium laptops, where variability is often pronounced. Looking ahead, these findings bear significant ramifications for both real-world implementations and prospective scholarly inquiry. The demonstrated efficacy of XGBoost in forecasting prices underscores the potency of ensemble learning methods in unraveling the intricate dynamics of technology valuation.

Advancements moving forward may prioritize refined feature engineering, meticulous hyper parameter tuning, and the integration of supplementary variables—such as brand perception and prevailing market tendencies—to augment predictive precision. These refinements could bolster the model's capability to generate highly accurate price estimations across diverse laptop categories. The insights derived from this investigation hold substantial value for key players within the laptop industry, spanning manufacturers, retailers, and consumers, by facilitating more strategic pricing decisions and well-informed purchasing choices. The approaches and conclusions delineated in this study establish a robust groundwork for crafting even more sophisticated price prediction frameworks, thereby fostering heightened market efficiency and enhancing overall consumer satisfaction within the competitive landscape of the laptop sector.

REFERENCES

1. Goswami, Shankha Shubhra, Rajesh Kumar Moharana, and Dhiren Kumar Behera. A new mcdm approach to solve a laptop selection problem. In Proceedings of Data Analytics and Management: ICDAM 2021, Volume 1, pp. 41-55. Singapore: Springer Nature Singapore, 2022. DOI: 10.1007/978-981-16-6289-8_5
2. Goswami, Shankha Shubhra, and Dhiren Kumar Behera. Best laptop model selection by applying integrated ahp-topsis methodology. International Journal of Project Management and Productivity Assessment (IJPMPA) 9, no. 2 (2021): 29-47. DOI: 10.4018/IJPMPA.2021070102
3. Aytaç Adalı, Esra, and Ayşegül Tuş Işık. The multi-objective decision-making methods based on MULTIMOORA and MOOSRA for the laptop selection problem. Journal of Industrial Engineering International 13 (2017): 229-237. DOI: https://doi.org/10.1007/s40092-016-0175-5
4. Sönmez Çakır, Fatma, and Mehmet Pekkaya. Determination of interaction between criteria and the criteria priorities in laptop selection problem. International Journal of Fuzzy Systems 22, no. 4 (2020): 1177-1190. DOI: https://doi.org/10.1007/s40815-020-00857-2
5. Lakshmi, T. Miranda, V. Prasanna Venkatesan, and A. Martin. Identification of a Better Laptop with Conflicting Criteria Using TOPSIS. International Journal of

Information Engineering & Electronic Business 7, no. 6 (2015). DOI: 10.5815/ijieeb.2015.06.05
6. Diana, Anita, and Achmad Solichin. Decision Support System with Fuzzy Multi-Attribute Decision Making (FMADM) and Simple Additive Weighting (SAW) In Laptop Vendor Selection. In 2020 Fifth International Conference on Informatics and Computing (ICIC), pp. 1-7. IEEE, 2020. DOI: 10.1109/ICIC50835.2020.9288587
7. Mitra, Soupayan, Shankha Shubhra Goswami, and MonayemParvej. Selection of the best laptop model by the application of fuzzy-AHP methodology. i-Manager's Journal on Management 14, no. 1 (2019): 33. DOI:10.26634/jmgt.14.1.16044
8. Raja, Chandrasekar, M. Ramachandran, Sathiyaraj Chinnasamy, and Sangeetha Rajkumar. A study on Laptop Computers Selection Problem Using the Grey Relational Analysis (GRA) Technique.
9. Aric, FransiskusAprilion, and Alexander Waworuntu. Android-based Decision Support System in Laptop Selection Using ELECTRE Method. In 2021 6th International Conference on New Media Studies (CONMEDIA), pp. 99-104. IEEE, 2021. DOI: 10.1109/CONMEDIA53104.2021.9617164
10. Fatma Sonmez Cakir, Mehmet Pekkaya. Determination of Interaction Between Criteria and the Criteria Priorities in Laptop Selection Problem. (2020). DOI: https://dx.doi.org/10.1007/s40815-020-00857-2
11. Harahap, Nur Hazimah Syani, and Afifah Zahraini. Laptop selection decision support system according to buyer criteria with the simple additive weighting method. Journal of Soft Computing Exploration 2, no. 2 (2021): 127-134. DOI: https://doi.org/10.52465/joscex.v2i2.49
12. Elsolia, Eslam. 2025. Laptop Price Prediction. Kaggle. Accessed February 6, 2025. https://www.kaggle.com/datasets/eslamelsolya/laptop-price-prediction/data.
13. Aalen, Odd O. A linear regression model for the analysis of life times. Statistics in medicine 8, no. 8 (1989): 907-925.
14. Yao, Weixin, and Longhai Li. A new regression model: modal linear regression. Scandinavian Journal of Statistics 41, no. 3 (2014): 656-671.
15. Poole, Michael A., and Patrick N. O'Farrell. The assumptions of the linear regression model. Transactions of the Institute of British Geographers (1971): 145-158.
16. Krämer, Walter, and Harald Sonnberger. The linear regression model under test. Springer Science & Business Media, 2012.
17. Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. John Wiley & Sons, 2021.
18. Maftoun, Mohammad, Nima Shadkam, Seyedeh Somayeh Salehi Komamardakhi, Zulkefli Mansor, and Javad HassannatajJoloudari. Malicious URL Detection using optimized Hist Gradient Boosting Classifier based on grid search method. arXiv preprint arXiv:2406.10286 (2024).
19. Tamim Kashifi, Mohammad, and Irfan Ahmad. Efficient histogram-based gradient boosting approach for accident severity prediction with multisource data. Transportation research record 2676, no. 6 (2022): 236-258.
20. Ayaru, Lakshmana, Petros-Pavlos Ypsilantis, Abigail Nanapragasam, Ryan Chang-Ho Choi, Anish Thillanathan, Lee Min-Ho, and Giovanni Montana. Prediction of outcome in acute lower gastrointestinal bleeding using gradient boosting. PLoS One 10, no. 7 (2015): e0132485.
21. Shi, Yu, Jian Li, and Zhize Li. Gradient boosting with piece-wise linear regression trees. arXiv preprint arXiv:1802.05640 (2018).
22. Persson, Caroline, Peder Bacher, Takahiro Shiga, and Henrik Madsen. Multi-site solar power forecasting using gradient boosted regression trees. Solar Energy 150 (2017): 423-436.
23. Zhang, Xinmeng, Chao Yan, Cheng Gao, Bradley A. Malin, and You Chen. Predicting missing values in medical data via XGBoost regression. Journal of healthcare informatics research 4 (2020): 383-394.
24. Dong, Jianwei, Yumin Chen, Bingyu Yao, Xiao Zhang, and Nianfeng Zeng. A neural network boosting regression model based on XGBoost. Applied Soft Computing 125 (2022): 109067.
25. Shehadeh, Ali, Odey Alshboul, Rabia Emhamed Al Mamlook, and Ola Hamedat. Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. Automation in Construction 129 (2021): 103827.
26. Pesantez-Narvaez, Jessica, Montserrat Guillen, and Manuela Alcañiz. Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. Risks 7, no. 2 (2019): 70.
27. Shahani, Niaz Muhammad, Xigui Zheng, Cancan Liu, Fawad Ul Hassan, and Peng Li. Developing an XGBoost regression model for predicting young's modulus of intact sedimentary rocks for the stability of surface and subsurface structures. Frontiers in Earth Science 9 (2021): 761990.