Predicting Smartphone Prices Based on Key Features Using Random Forest and Gradient Boosting Algorithms in a Data Mining Framework

Retno Wahyusari[1,*], Zahara Nabila[2]

[1,2]*Informatics Department, Ronggolawe College of Technology, Tuban, Indonesia*

(Received: January 8, 2025; Revised: February 15, 2025; Accepted: May 22, 2025; Available online: July 23, 2025)

Abstract

This study aims to predict smartphone prices using machine learning models, specifically Random Forest and Gradient Boosting algorithms, based on various smartphone features such as internal memory, RAM, processor speed, battery capacity, and camera specifications. The dataset, consisting of 980 smartphones available in India, was preprocessed to handle missing values and categorical variables, ensuring it was ready for model training. The models were evaluated using Mean Squared Error (MSE) and R-squared (R^2) scores, with Gradient Boosting outperforming Random Forest in terms of predictive accuracy. Key findings from the feature importance analysis revealed that internal memory, RAM, and processor speed were the most influential features in determining smartphone prices. The results indicate that machine learning models, particularly tree-based algorithms, are effective tools for predicting smartphone prices based on hardware specifications. This study has practical implications for businesses and consumers, as it provides insights into the factors influencing smartphone prices, helping businesses optimize pricing strategies and assisting consumers in making more informed purchasing decisions. Future research could explore deep learning models and incorporate additional features, such as market demand and consumer sentiment, to improve prediction accuracy.

Keywords: Feature Importance, Gradient Boosting, Price Prediction, Random Forest, Smartphone

1. Introduction

The significance of smartphones in contemporary society is continuously escalating, influenced by their integration into various aspects of daily life. As technology adapts and evolves, smartphones have transcended their traditional roles, emerging as essential tools for communication, information access, entertainment, and education. This evolution has been particularly pronounced during the COVID-19 pandemic, where the necessity for remote connectivity became paramount. Sela et al highlight that smartphones have become pivotal for maintaining social connections during extended periods of isolation, underscoring their indispensable role in modern life amid crises [1].

The shift towards smartphones as primary computing devices has been remarkable; they now incorporate functionalities that surpass those of traditional desktop computers. Riadi et al note that smartphones facilitate global information exchange, making it easier for people to share information and access various applications that enhance user experience and accessibility [2]. This transition is exacerbated by advancements in technology, where mobile applications are increasingly developed for diverse fields such as healthcare, education, and leisure, as described by [3]. These devices have transformed into multifunctional tools that encapsulate the capabilities of numerous devices within a single, portable format.

Moreover, the economic implications of this evolution cannot be understated. The smartphone market exemplifies a highly competitive environment characterized by rapid innovation cycles and shifting consumer preferences. To maintain competitive advantages, manufacturers continuously refine their product offerings and pricing strategies. Frintika and Rachmawati emphasize that product pricing significantly impacts purchase intentions among consumers, particularly in emerging markets like Indonesia [4]. This observation is echoed by Chandiona et al, who illustrate that during the COVID-19 pandemic, pricing, alongside brand image and product features, affected young consumers'

***Corresponding author: Retno Wahyusari (retnowahyusari@gmail.com)**
DOI: https://doi.org/10.47738/ijaim.v5i2.100

intentions to purchase smartphones [5]. The increasing reliance on smartphones has also coincided with the expansion of mobile applications that further enhance user engagement and productivity. The mobile market is witnessing significant growth as applications cater to virtually every consumer need, as articulated by Wohllebe et al, who underscore that consumers utilize apps across diverse life scenarios, from travel planning to health management [3]. This trend not only reinforces the importance of smartphones but also indicates the need for continuous research into network performance and the implications of mobile traffic, as noted by [6].

Predicting smartphone prices accurately is a multifaceted challenge in an industry that is characterized by rapid technological advancements and varying consumer preferences. The complexity arises from numerous technical features that smartphones offer, each with the potential to affect market prices differently. Given the intricate interplay of these features such as brand, design, specifications, and market dynamics, leveraging machine learning algorithms has emerged as a promising approach to enhance the accuracy of smartphone price predictions. Zhao's study highlights this by indicating that advanced machine learning methods can effectively analyze the multitude of features that define smartphone pricing amidst the constantly fluctuating market conditions [7].

One of the significant obstacles in predicting smartphone prices lies in the sheer volume and variety of features present in new models. As new smartphones flood the market, each embodying unique specifications, accurately determining pricing metrics becomes increasingly complex. Ercan and Şimşek's examination of machine learning models reveals a similar challenge, advocating for comprehensive datasets that encompass a wide array of features in order to train predictive models adeptly. They report that utilizing different algorithms can yield significant accuracy improvements, with Support Vector Machine achieving notable success in categorizing smartphone prices based on these intricate feature sets [8]. Moreover, integrating diverse models ensures the mitigation of overfitting while enhancing predictive performance, making it a vital consideration in price prediction models.

Despite the advancements brought by machine learning algorithms, the reliance on historical datasets can often encapsulate biases based on trends within the tech market. However, drawing from frameworks used in real estate price predictions, such as those discussed by Gawali in context with home prices, can provide foundational methodologies applicable to consumer electronics. The relevance of exploratory data analysis in identifying key correlations between features and price points can significantly enhance the prediction models employed in smartphone pricing scenarios [9]. Employing a systematic approach that involves identifying variables, cleaning datasets, and assessing outliers can facilitate the development of robust predictive models that adapt to market fluctuations.

Additionally, it is evident that the influence of economic factors cannot be overlooked. Changes in global supply chains, trade policies, and economic conditions invariably affect production costs and consumer affordability. The need for flexible pricing strategies that can accommodate economic shifts is paramount. The growing importance of adaptable pricing mechanisms is corroborated by the predictive trends noted in various studies, suggesting that dynamic pricing models—which account for external market fluctuations—might offer more accurate pricing forecasts than static models alone.

Moreover, the competitive landscape of the smartphone market necessitates continuous innovation in predictive methodologies. The focus on deep learning and more advanced machine learning applications has shown promise in enhancing predictive capabilities. Machine learning’s ability to recognize patterns across vast datasets allows for more nuanced price predictions that account for intricate relationships among product features, consumer behavior, and market conditions. High-performing algorithms can thus provide essential insights for manufacturers, helping streamline production and pricing strategies in response to forecasting results [7], [10].

The aim of this research is to leverage data mining algorithms to predict smartphone prices based on multiple technical specifications and features. The rapid evolution of the smartphone industry presents both opportunities and challenges, particularly in terms of pricing dynamics motivated by the complex interplay of numerous factors, including technological advancements, consumer preferences, and competitive market practices. With the proliferation of smartphones featuring various attributes—such as processing power, memory capacity, design aesthetics, and brand reputation—there is an increasing need for sophisticated predictive models that can accurately align smartphone pricing with consumer expectations and market trends.

Moreover, research conducted by Frintika and Rachmawati suggests that a mix of variables—including product features, brand image, and pricing strategies—significantly influences consumer purchasing intentions [4]. This indicates the importance of not only understanding the technical attributes but also integrating marketing perspectives into price prediction models. By acknowledging the multifaceted nature of price determination, this research aims to bridge the gap between feature specifications and perceived consumer value, ultimately enhancing the reliability of pricing projections.

The scope and significance of this research lie in its potential to provide valuable insights into smartphone pricing trends, which are critical for businesses, marketers, and consumers. For businesses, understanding the factors influencing smartphone prices can guide inventory decisions, product launches, and promotional strategies. Marketers can benefit from predicting price fluctuations to tailor advertising campaigns and target the right consumer segments. Additionally, consumers can leverage this research to make informed purchasing decisions, potentially saving money by identifying the most cost-effective smartphone options based on their desired features.

This study contributes to the field of price prediction by applying data mining techniques, specifically Random Forest and Gradient Boosting algorithms, to model smartphone pricing. The findings offer a deeper understanding of which features most significantly affect smartphone prices, enriching current price prediction models. The practical application of these algorithms allows businesses and consumers to better anticipate price trends and make data-driven decisions. Moreover, this research highlights the effectiveness of machine learning techniques in real-world applications, demonstrating how data mining can optimize pricing strategies and improve market forecasting.

2. Literature Review

2.1. Overview of Price Prediction Models

Price prediction models for consumer electronics, particularly smartphones, have garnered substantial interest in recent years due to the rapid pace of technological advancements and the growing complexity of pricing strategies. This overview examines existing literature on various models employed to predict smartphone prices, encompassing machine learning algorithms, statistical analyses, and hybrid approaches. By synthesizing insights from multiple studies, one can better understand the methodologies utilized and their effectiveness in accurately forecasting prices. In the realm of smartphone price prediction, numerous machine learning algorithms have emerged as prominent tools. Zhao [7] elaborates on various techniques, including Decision Tree Regression, Support Vector Regression (SVR), and Random Forest Regression, highlighting their effectiveness in associating smartphone specifications—such as processor capabilities, memory, and camera quality—with market prices. Through comprehensive data analysis, these models harness past observations to identify meaningful patterns that dictate price fluctuations, underscoring machine learning's capacity for making data-driven predictions in a dynamic market.

Tarigan [11] conducts an in-depth exploration using Conjoint Analysis to establish optimal pricing strategies for new smartphones. His findings suggest that manufacturers adopting skimming pricing strategies, based on a product's features and perceived value, tend to enhance their competitive edge significantly. This approach demonstrates how consumer preferences and feature importance can be effectively correlated, thereby facilitating more accurate price predictions aligned with market expectations. The utilization of K-Nearest Neighbors (KNN) and Linear Regression models has also been outlined in the literature. In Chen’s study [12], both techniques are employed to predict smartphone prices, with KNN exhibiting a slight edge in terms of accuracy. This indicates that simpler models can still provide reliable forecasts, especially when dealing with less complex datasets. The study emphasizes the value of model selection in enhancing prediction performance, advocating for methodological pluralism in this field.

Moreover, Li's research [13] investigates the comparative efficacy of Decision Trees and SVR specifically for smartphone price prediction, showcasing the nuanced capabilities of these algorithms in managing complex relationships within data. Both models offer unique strengths, with Decision Trees providing intuitive classification while SVR adeptly handles regression tasks. Such comparative studies illuminate the importance of model choice based on the specifics of the dataset and desired outcomes. Gaining perspective from adjacent fields can also enhance the understanding of price prediction methodologies.

2.2. Feature Selection in Price Prediction

Feature selection is a crucial aspect of developing effective predictive models for smartphone prices. The features chosen for analysis can significantly influence model accuracy and the insights drawn from the predictive outcomes. A review of existing literature indicates that a wide range of features is commonly used in predicting smartphone prices, reflecting the multifaceted nature of consumer electronics. One of the primary features utilized across several studies is camera quality. According to Frintika and Rachmawati, camera specifications such as megapixels and aperture size have a significant impact on consumer purchasing decisions for smartphones, confirming their importance in pricing models [4]. Consumers increasingly demand high-quality photography capabilities in smartphones, which drives manufacturers to include advanced camera systems that can substantially influence price determinations.

Battery capacity is another frequently analyzed feature. Research by Haverila indicates that this attribute is particularly important to certain demographics, such as male users who prioritize battery life alongside hardware quality and display size [14]. Efficient battery performance is critical for smartphones, and as manufacturers incorporate advanced battery technologies, this feature serves as a significant determinant of device cost. The memory capacity of smartphones, encompassing both RAM and internal storage, is also a pivotal feature in pricing models. A higher RAM capacity often translates to better multitasking capabilities and smoother performance, which are highly valued by consumers in today's fast-paced digital environment. High storage capacity facilitates larger app installations and data storage needs, making smartphones with ample memory more expensive. Tarigan discusses the significance of combining relevant features to develop tailored pricing strategies that can enhance overall profitability [11].

Processor specifications play a vital role in price prediction models. The performance level of a smartphone is heavily influenced by its processor type, clock speed, and the number of cores. Consumers, especially tech-savvy individuals, often consider these specifications when evaluating smartphones, which significantly affects their purchasing decisions. Ercan and Şimşek emphasize that understanding the relationship between processing power and consumer preferences can help effectively categorize smartphone prices [8]. Display technology and size is another common feature that influences pricing in the smartphone market. The transition from traditional LCD screens to OLED and other advanced display technologies has led to significant price variation in smartphones. As noted by Rakib et al, consumers often seek devices that provide superior visual experiences, which can greatly impact their willingness to pay higher prices for certain models [15]. Thus, considering display quality and size as features in price prediction models is essential for gaining a comprehensive understanding of market dynamics.

2.3. Key Studies Applying Random Forest and Gradient Boosting in Price Prediction

The application of Random Forest and Gradient Boosting algorithms has become increasingly prominent in various domains for predicting prices, showcasing their efficacy across different industries. Research [16] discusses the application of Random Forest in predicting diamond prices, contrasting it with linear regression and decision tree models. The study highlights that Random Forest effectively mitigates the overfitting problem typically associated with decision trees by aggregating multiple sub-decision trees, leading to more reliable price predictions [16]. Study [17] investigates how Random Forest can be utilized for predicting taxi fare based on trip distance using regression analysis. Their findings demonstrate that Random Forest outperforms other models in accuracy, achieving the lowest RMSE, indicating its effectiveness for prediction tasks involving complex, non-linear relationships [17]. Research [18] applies Random Forest to predict clean energy stock prices, revealing high accuracy levels for price forecasting over a 20-day horizon. This study underscores Random Forest's strength in producing reliable price predictions in the financial sector compared to traditional models [18].

Study [19] applies Gradient Boosting for water demand forecasting, analyzing the impact of various factors on consumption. The study illustrates how Gradient Boosting can improve predictions by effectively handling non-linear relationships in the data, indicating its applicability in resource management scenarios [19]. Research [20] investigates rent price prediction leveraging advanced machine learning methods, including Gradient Boosting and Random Forest. Results demonstrate that Gradient Boosting outperforms other models in predictive accuracy, showcasing its effectiveness in real estate applications [20]. Study [21] employs Gradient Boosting Regression to predict hotel revenue, comparing it with Support Vector Regression. The study confirms the superiority of Gradient Boosting in

managing intricate relationships within the data, thus providing more accurate revenue predictions for the hospitality industry [21].

2.4. Formulae and Methods Used

In the context of price prediction, particularly when using data mining algorithms like Random Forest and Gradient Boosting, certain performance metrics are essential for evaluating the efficacy of the models employed. The following equations outline two commonly utilized metrics in regression analysis: MSE and R^2. The MSE measures the average squared difference between the actual and predicted values. It is given by the equation:

$$MSE = \frac{1}{n}\sum i = 1^n (yi - \hat{y}_i)^2 \quad (1)$$

Where n is the total number of observations, yi represents the actual value, $\hat{y}_i$ denotes the predicted value from the model. MSE is a crucial metric because it provides a measure of how well the model predicts the target variable, with lower values indicating better predictive accuracy. In practical applications, such as smartphone price prediction, MSE can help determine how closely the model's outputs match actual market prices. The use of MSE was highlighted by Agusdin et al, who discussed its significance in forecasting accuracy assessments [22]. The R^2 value quantifies the proportion of variance in the dependent variable that can be explained by independent variables in the model. It is represented by the equation:

$$R^2 = 1 - \frac{\sum i = 1^n (yi - \hat{y}i)^2}{\sum i = 1^n (y_i - \bar{y})^2} \quad (2)$$

Where yi is the actual value, $\hat{y}i$ is the predicted value, and $\bar{y}$ is the mean of the actual values. R^2 serves as an essential statistic indicating the goodness-of-fit of the model, where values closer to 1 signify a greater proportion of explanatory variance captured by the model. The relevance of R^2 in model assessment has been employed to gauge the performance of predictive models effectively [23], [24] .

These equations have been widely adopted across studies that analyze price prediction models. Nwanze et al employed various statistical indices, including R2 and RMSE, to evaluate their developed solar radiation forecasting models, which shows the importance of these metrics in appraising model performance [25]. Tarun and Sriramya highlighted the use of MSE and RMSE as critical metrics in their analysis of taxi fare predictions using Random Forest and regression methodologies, emphasizing how these metrics convey the predictive accuracy of the models utilized [17]. Zhang et al, in their study on second-hand sailboat pricing using Random Forest, indicated the significance of MSE alongside regression models to ascertain prediction accuracy [26].

3. Methodology

3.1. Data Loading and Initial Inspection

The first step in the process is loading the dataset from Kaggle. The `filepath` parameter specifies the location of the dataset file. Upon loading the data, the column names are standardized by stripping any leading or trailing spaces and converting them to lowercase to prevent any case sensitivity issues during analysis. The target column, `price`, is then cleaned to ensure it only contains numeric values. This is done by removing any non-numeric characters using regular expressions and converting the column to a numeric type. Any invalid or non-numeric entries are handled by converting them to `NaN` and subsequently removing rows with missing price values. The dataset is then inspected by displaying its structure and summary statistics, checking for missing values, and ensuring the data is correctly formatted and ready for analysis.

3.2. EDA

EDA is conducted to uncover patterns and relationships within the dataset. First, a correlation matrix is computed to assess the relationships between numerical features and their correlation with the target variable, `price`. The matrix helps identify which features have the strongest linear correlation with the price. This is visualized through a heatmap, where higher correlations are represented with more intense colors. The EDA also includes several key visualizations: a bar plot showing the average price by brand, and scatter plots illustrating the relationship between price and other

important features like RAM and battery capacity. These visualizations help to understand how these features influence smartphone prices and provide insights into the distribution of data.

3.3. Data Preprocessing for Modeling

Data preprocessing is an essential step to ensure that the dataset is ready for machine learning algorithms. Missing values in the numerical columns are imputed using the median, while missing values in categorical columns are filled with the mode (most frequent value). This ensures that there are no gaps in the data that could interfere with model training. Categorical variables, such as `brand_name` and `os`, are transformed into numerical format using one-hot encoding, which creates binary columns for each category. The model-specific column, which could lead to overfitting, is dropped to ensure that only meaningful features are used for prediction. After these transformations, the dataset is inspected again to confirm that there are no remaining missing values and that it is ready for use in training the models.

3.4. Model Development and Evaluation

In the model development phase, two machine learning algorithms are selected: Random Forest Regressor and Gradient Boosting Regressor. These algorithms are chosen for their ability to handle complex relationships between features and target variables. The data is split into training and testing sets using an 80/20 split, ensuring that the model is trained on a majority of the data and tested on unseen data to evaluate performance. For both models, key parameters such as `n_estimators=100` (the number of trees or boosting stages) and `random_state=42` (to ensure reproducibility) are defined. Random Forest also uses `n_jobs=-1` to leverage all CPU cores during training, improving efficiency. Gradient Boosting utilizes `learning_rate=0.1` to control the model's step size and `max_depth=3` to prevent overfitting. The models are trained on the training data and evaluated using MSE and R^2 metrics. MSE measures the average squared differences between the predicted and actual values, while R^2 indicates how well the model explains the variance in the target variable.

3.5. Results Visualization

After training and evaluating the models, results are visualized to better understand model performance. Scatter plots are used to compare the actual vs. predicted smartphone prices, with a reference line representing perfect predictions. This helps to visually assess how close the predicted values are to the actual values. Additionally, feature importance is visualized to show which features contributed the most to the model’s predictions. This is particularly useful for interpreting which smartphone characteristics (e.g., RAM, battery capacity, brand) have the largest impact on price prediction. Bar charts are generated to visualize the relative importance of the top features, with the most influential features listed first. These visualizations aid in understanding the model's behavior and provide insights into which factors are most important for price prediction.

3.6. Model Checkpointing

Once the models are trained and evaluated, they are saved for future use. The models are serialized and stored using joblib, allowing them to be easily reloaded without the need for retraining. The `save_models_checkpoint` function specifies the `directory` where the models will be saved, and `filename` defines the naming convention used for the saved models. The models are stored in a designated directory, ensuring they can be reloaded for future predictions or further analysis. This step is important for maintaining an efficient workflow, as it eliminates the need to retrain the models each time they are needed. The saved models can be reused as part of a production system or for testing with new data.

4. Results and Discussion

4.1. Results from Initial Inspection

The dataset was successfully loaded, containing 980 smartphone entries across 26 features. These features provide comprehensive details about the smartphones, including attributes such as brand, model, price, rating, processor type, RAM, camera specifications, and screen size. After loading the data, the column names were standardized by stripping any whitespace and converting them to lowercase. This ensured that there would be no issues with case sensitivity during analysis and subsequent modeling steps. The target column, `price`, was cleaned by removing non-numeric

characters, ensuring that the values were consistent and could be used for regression. The column was converted to a numeric format, with any invalid or non-numeric entries removed from the dataset.

An initial inspection of the dataset was conducted using methods like `head()`, `info()`, and `describe()`. This revealed that the dataset contained 980 rows and 26 columns. Some columns had missing values, particularly the `rating`, `processor_brand`, and `fast_charging` columns. The `rating` column had 101 missing values, while the `processor_brand` column had 20 missing entries. Other columns, such as `fast_charging` and `num_cores`, also had a small number of missing values. Missing values were handled appropriately by imputing numerical columns with their median values and categorical columns with the mode. This ensured the dataset was complete and could be used for model training without any issues related to missing data.

4.2. EDA and Data Preprocessing Finding

EDA was performed to uncover patterns and relationships in the dataset. The first step was generating a correlation matrix for the numerical features to understand how they relate to the target variable, `price` (figure 1). This matrix revealed several features with strong positive correlations with price. The top correlations were with `internal_memory` (0.557), `processor_speed` (0.474), and `ram_capacity` (0.386), indicating that smartphones with more internal memory, faster processors, and higher RAM are generally priced higher. The correlation heatmap generated from this matrix visually confirmed these findings, with higher correlations highlighted in darker shades. Further, the correlation between `rating` and `price` was found to be moderately positive (0.283), suggesting that higher-rated smartphones tend to have higher prices. Other features like `screen_size`, `resolution_height`, and `battery_capacity` also showed moderate correlations with the target variable.

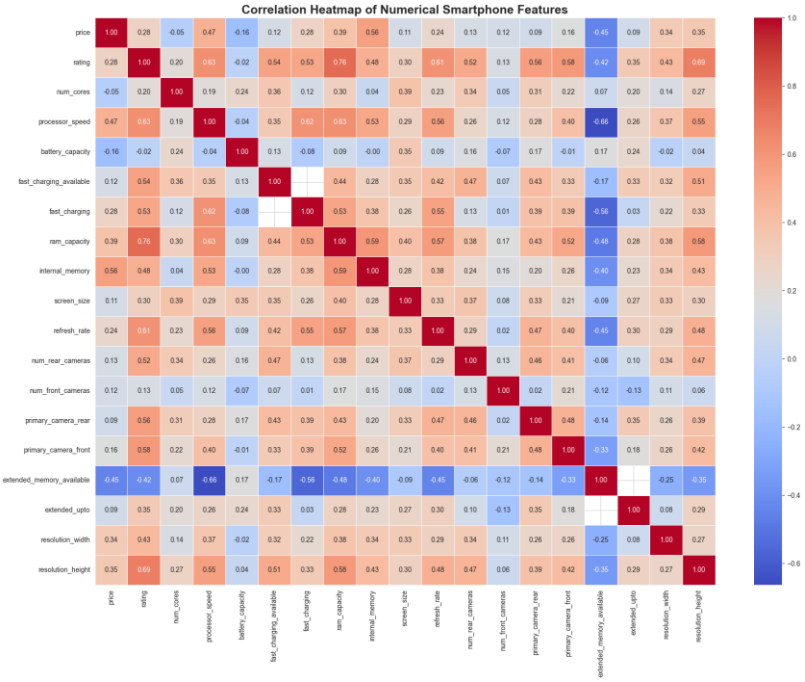

Figure 1. Correlation Matrix

To further explore the relationships, several key visualizations were created. These included a bar plot illustrating the average price by brand (figure 2), which showed that premium brands like Apple and Samsung tended to have higher average prices compared to brands like Xiaomi or Realme and also scatter plots were created to analyze the relationship between price and features such as `battery_capacity` (figure 3). These plots revealed that smartphones with higher RAM and larger battery capacities tended to have higher prices, supporting the results from the correlation matrix.

 ISSN 2776-8007

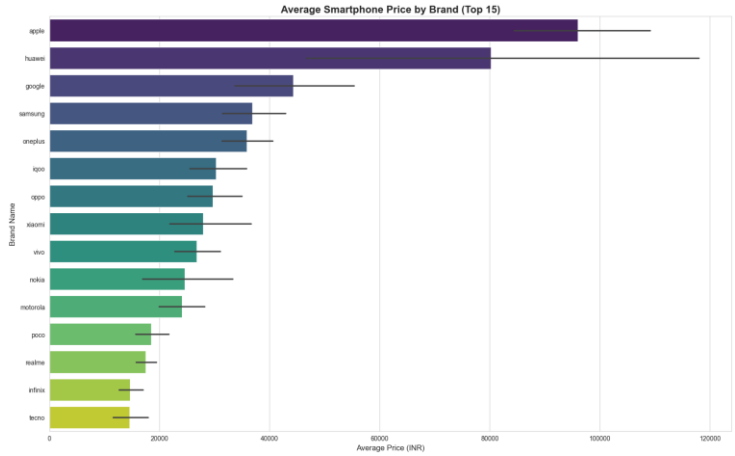

Figure 2. Average Smartphone Price by Brand (Top 15)

Figure 3. Scatter Plot of Price vs Battery Capacity

The dataset underwent a thorough preprocessing phase to prepare it for machine learning model development. Missing values were imputed for both numerical and categorical columns. For numerical columns like `rating`, `num_cores`, and `battery_capacity`, the median value was used to impute missing data. For categorical columns like `processor_brand` and `os`, the most frequent value (mode) was used for imputation. After imputation, no missing values remained in the dataset. Categorical variables such as `brand_name`, `processor_brand`, and `os` were encoded using one-hot encoding. This transformation created binary columns for each unique category in these features, making them suitable for machine learning algorithms. The `model` column, which contained specific model names, was dropped to prevent overfitting, as it was too detailed and likely to introduce noise into the model. The final dataset after preprocessing had 980 rows and 81 features, with no missing values and all categorical variables encoded.

4.3. Model Development and Evaluation Results

Two machine learning models were trained: Random Forest Regressor and Gradient Boosting Regressor. The data was split into training and testing sets, with 784 samples used for training (80%) and 196 samples reserved for testing (20%). Both models were then trained using the training data, and their performance was evaluated using two key metrics: MSE and R^2.

The Random Forest model was trained with 100 trees (`n_estimators=100`) and the random state was set to 42 to ensure reproducibility. The model achieved an MSE of 199,004,693.49, which indicates the average squared difference between the actual and predicted prices. The R^2 score for the Random Forest model was 0.7922, meaning that the model explained approximately 79.22% of the variance in the price data.

The Gradient Boosting model was trained with 100 boosting stages (`n_estimators=100`), a learning rate of 0.1, and a maximum depth of 3 for the individual trees (`max_depth=3`). The model performed slightly better than Random

 ISSN 2776-8007

Forest, achieving an MSE of 165,030,523.72 and an R^2 score of 0.8277. This indicates that Gradient Boosting was able to explain approximately 82.77% of the variance in the price data, with a lower MSE, suggesting better predictive accuracy.

4.4. Visualization

The performance of the models was visualized through several plots. Scatter plots comparing actual vs. predicted prices for both models were generated (figure 4 and figure 5), with a 45-degree reference line representing perfect predictions. These plots revealed that both models provided reasonably accurate predictions, although there were some discrepancies, particularly in the higher price range where the models tended to slightly under-predict the actual prices.

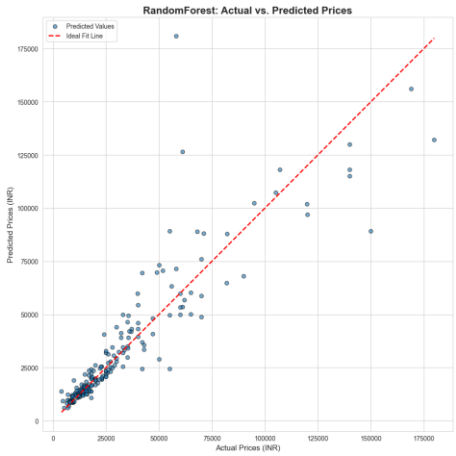

Figure 4. Actual vs Predicted Prices Graph for Random Forest

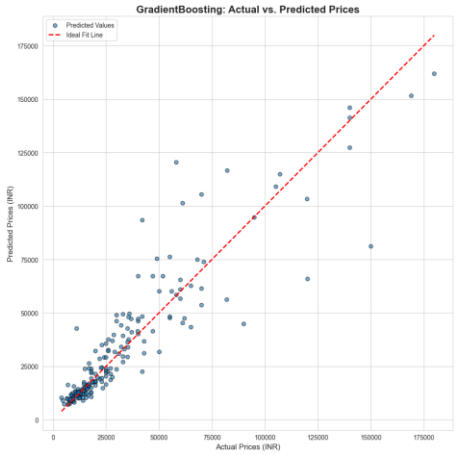

Figure 5. Actual vs Predicted Prices Graph for GradientBoosting

Additionally, feature importance was visualized for both models (figure 6 and figure 7). The importance of each feature in predicting smartphone prices was plotted, with the most influential features, such as `internal_memory`, `ram_capacity`, and `processor_speed`, showing up with the highest importance scores. This visualization helped to interpret which features had the most significant impact on the predictions, reinforcing the insights obtained during the EDA phase. Overall, the visualizations provided a clear picture of the models' performance and the importance of various features in predicting smartphone prices.

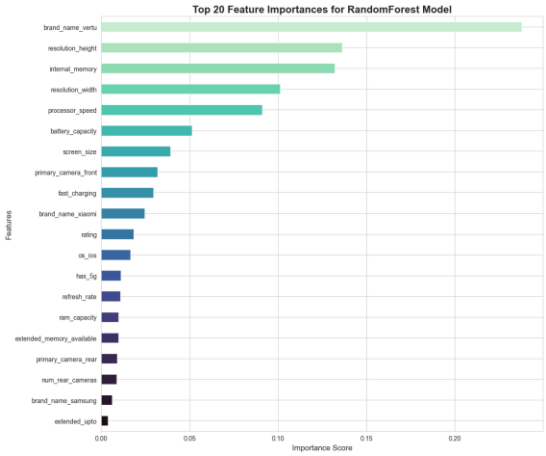

Figure 6. Top 20 Feature Importance for Random Forest

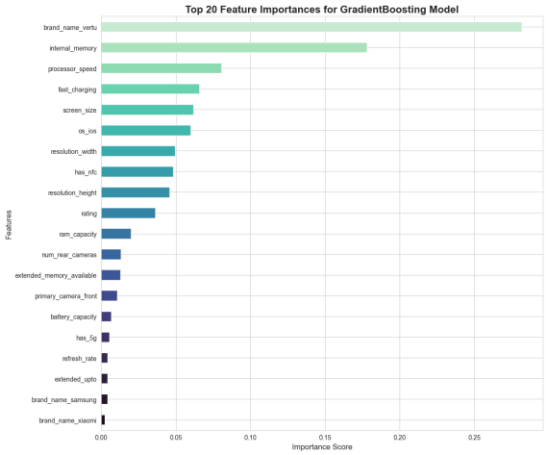

Figure 7. Top 20 Feature Importance for Gradient Boosting

4.5. Discussion

The results of this study indicate that both Random Forest and Gradient Boosting models performed well in predicting smartphone prices, with Gradient Boosting achieving slightly better performance in terms of R^2 and MSE. This finding aligns with previous studies, which have shown that tree-based models like Random Forest and Gradient Boosting are effective in capturing complex, non-linear relationships between features and target variables. These models can handle a large number of input features and interactions, making them suitable for predicting prices in the smartphone market. The performance metrics, particularly the R^2 scores (0.7922 for Random Forest and 0.8277 for Gradient Boosting), suggest that both models were able to explain a significant portion of the variance in smartphone prices. These results are consistent with other research on price prediction, where tree-based models have been successful in achieving high accuracy.

Key insights from the feature importance analysis revealed that internal memory, ram capacity, and processor speed were the most influential features in predicting smartphone prices. These features were highly correlated with price in the correlation matrix and played a significant role in both the Random Forest and Gradient Boosting models. The strong correlation between internal memory and price is expected, as smartphones with larger storage capacities tend to be priced higher. Similarly, higher RAM and faster processors contribute to a better user experience, making them more desirable and thus influencing the price. Other features like battery capacity and camera specifications also had moderate importance, which aligns with consumer preferences for longer battery life and better camera quality in smartphones. These findings emphasize the importance of hardware specifications in determining the price of smartphones, as has been observed in previous studies on smartphone pricing.

Both Random Forest and Gradient Boosting models have their strengths and weaknesses. Random Forest is a robust model that can handle large datasets and provides good performance even without extensive parameter tuning. Its strength lies in its ability to reduce overfitting by averaging the predictions of multiple trees. However, it can be computationally expensive when the number of trees increases, and its interpretability can be limited due to the ensemble nature of the model. On the other hand, Gradient Boosting tends to perform better in terms of accuracy, as it builds trees sequentially, correcting the errors of previous trees. However, it is more sensitive to overfitting and requires careful tuning of parameters like the learning rate and tree depth to avoid this. Despite these differences, both models demonstrated strong predictive accuracy, and their complementary strengths could be leveraged in practice depending on the specific use case, with Gradient Boosting being slightly preferable for better predictive performance.

5. Conclusion

In this study, the effectiveness of Random Forest and Gradient Boosting models in predicting smartphone prices was evaluated. Both models demonstrated strong performance, with Gradient Boosting slightly outperforming Random Forest in terms of predictive accuracy. The models were able to explain a substantial portion of the variance in smartphone prices, achieving R^2 scores of 0.7922 and 0.8277, respectively. The most influential features in predicting prices were found to be internal memory, ram capacity, and processor speed, which had the highest impact on the model's predictions. These findings confirm that hardware specifications are crucial factors in determining smartphone prices, supporting similar results found in previous research on price prediction.

The practical implications of this study are significant for both businesses and consumers. For businesses, the ability to predict smartphone prices accurately can guide inventory management, pricing strategies, and marketing efforts. By understanding which features most influence price, companies can make informed decisions on product offerings and target pricing. For consumers, the study offers valuable insights into the factors that drive smartphone prices, helping them make more informed purchasing decisions based on their preferences for certain features, such as storage capacity or processor speed, without overpaying for unnecessary specifications.

However, there are some limitations to this study. One key limitation is the dataset's constraints, as it only includes smartphones available in India, potentially limiting its generalizability to other markets. Additionally, while the models performed well, there could be biases inherent in the data, such as price inflation for premium brands, that might affect the predictions. Future research could address these limitations by incorporating a broader range of smartphone data from various regions. Additionally, applying deep learning models could further improve prediction accuracy by capturing more complex relationships between features. Future work could also explore the inclusion of more granular features, such as market demand trends or consumer sentiment, to enhance the robustness of the predictions.

6. Declarations

6.1. Author Contributions

Conceptualization: R.W., Z.N.; Methodology: R.W., Z.N.; Software: R.W.; Validation: Z.N.; Formal Analysis: R.W.; Investigation: R.W.; Resources: Z.N.; Data Curation: R.W.; Writing – Original Draft Preparation: R.W.; Writing – Review and Editing: R.W., Z.N.; Visualization: R.W.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Sela, N. Rozenboim, and H. C. Ben-Gal, "Smartphone Use Behavior and Quality of Life: What Is the Role of Awareness?," *Plos One*, 2022, doi: 10.1371/journal.pone.0260637.

[2] I. Riadi, A. Yudhana, G. Pramuja, and I. Fanani, "Mobile Forensic Tools for Digital Crime Investigation: Comparison and Evaluation," *International Journal of Safety and Security Engineering*, 2023, doi: 10.18280/ijsse.130102.

[3] A. Wohllebe, P. Dirrler, and S. Podruzsik, "Mobile Apps in Retail: Determinants of Consumer Acceptance – A Systematic Review," *International Journal of Interactive Mobile Technologies (Ijim)*, 2020, doi: 10.3991/ijim.v14i20.18273.

[4] M. T. Frintika and I. Rachmawati, "The Influence of E-Wom, Brand Image, Product Features, and Product Price on Purchase Intention for the Samsung Galaxy S23 Smartphone in Indonesia," *Manajemen Dan Kewirausahaan*, 2023, doi: 10.53682/mk.v4i2.7651.

[5] D. F. Chandiona, S. K. Kallier, and K. Makhitha, "Determinants Affecting Young Consumers' Smartphone Purchase Intention During Covid-19 Pandemic," *Mediterranean Journal of Social Sciences*, 2024, doi: 10.36941/mjss-2024-0002.

[6] S. Zhang, Z. Zhao, H. Guan, and H. Yang, "A Modified Poisson Distribution for Smartphone Background Traffic in Cellular Networks," *International Journal of Communication Systems*, 2016, doi: 10.1002/dac.3117.

[7] Z. Zhao, "Predicting Smartphone Prices Using Machine Learning Algorithms," *Applied and Computational Engineering*, 2024, doi: 10.54254/2755-2721/95/20241765.

[8] S. İ. AKSOY ERCAN and M. Şimşek, "Mobile Phone Price Classification Using Machine Learning," *International Journal of Advanced Natural Sciences and Engineering Researches*, 2023, doi: 10.59287/ijanser.791.

[9] Prof. P. Gawali, "House Price Prediction Using R Language," *Interantional Journal of Scientific Research in Engineering and Management*, 2023, doi: 10.55041/ijsrem26175.

[10] Berlilana and A. M. Wahid, "Time Series Analysis of Bitcoin Prices Using ARIMA and LSTM for Trend Prediction," *Journal of Digital Market and Digital Currency*, vol. 1, no. 1, Art. no. 1, May 2024, doi: 10.47738/jdmdc.v1i1.1.

[11] S. Tarigan, "Using Conjoint Analysis to Predict the Launch Price of a New Smartphone," *Journal of Management Studies and Development*, 2023, doi: 10.56741/jmsd.v2i01.221.

[12] Y. Chen, "Prediction of Different Types of Mobile Phone Prices Based on Machine Learning Models," *Highlights in Science Engineering and Technology*, 2024, doi: 10.54097/shgcew53.

[13] X. H. Li, "Smartphone Price Prediction Using Decision Tree and Support Vector Regression (SVR)," *Applied and Computational Engineering*, 2024, doi: 10.54254/2755-2721/2025.18475.

[14] C. Chang, W.-X. Zhao, and J. Hajiyev, "An Integrated Smartphone and Tariff Plan Selection for Taxi Service Operators: McDm and RStudio Approach," *Ieee Access*, 2019, doi: 10.1109/access.2019.2903201.

[15] Md. R. Hafiz Rakib, S. A. Kabir Pramanik, Md. A. Amran, Md. N. Islam, and Md. O. Faruk Sarker, "Factors Affecting Young Customers' Smartphone Purchase Intention During Covid-19 Pandemic," *Heliyon*, 2022, doi: 10.1016/j.heliyon.2022.e10599.

[16] Z. Ouyang, "Research on the Diamond Price Prediction Based on Linear Regression, Decision Tree and Random Forest," *Highlights in Business Economics and Management*, 2024, doi: 10.54097/13ccwv59.

[17] G. V. Sai Tarun and P. Sriramya, "Analyzing Ola Data for Predicting Price Based Trip Distance Using Random Forest and Linear Regression Analysis," 2022, doi: 10.3233/apc220086.

[18] P. Sadorsky, "A Random Forests Approach to Predicting Clean Energy Stock Prices," *Journal of Risk and Financial Management*, 2021, doi: 10.3390/jrfm14020048.

[19] M. Xenochristou, C. W. Hutton, J. Hofman, and Z. Kapelan, "Water Demand Forecasting Accuracy and Influencing Factors at Different Spatial Scales Using a Gradient Boosting Machine," *Water Resources Research*, 2020, doi: 10.1029/2019wr026304.

[20] H. R. Chan, "Rent Price Prediction With Advanced Machine Learning Methods: A Comparison of California and Texas," *Highlights in Science Engineering and Technology*, 2024, doi: 10.54097/84vvv580.

[21] M. Y. Anshori, "Predicting Hotel Revenue Using Gradient Boosting Regression and Support Vector Regression: A Comparative Analysis," 2025, doi: 10.21203/rs.3.rs-6910156/v1.

[22] R. P. Agusdin, S. P. Tahalea, and V. A. Permadi, "Forecasting the Poverty Rates Using Holt's Exponential Smoothing," *Matrik Jurnal Manajemen Teknik Informatika Dan Rekayasa Komputer*, 2024, doi: 10.30812/matrik.v23i2.2672.

[23] A. L. Chaves Gurgel *et al.*, "Mathematical Models to Predict Dry Matter Intake and Milk Production by Dairy Cows Managed Under Tropical Conditions," *Agriculture*, 2023, doi: 10.3390/agriculture13071446.

[24] M. Murakami *et al.*, "Improved Formula for Predicting Hemodialyzability of Intravenous and Oral Drugs," *Blood Purification*, 2021, doi: 10.1159/000513152.

[25] N. E. Nwanze, S. C. Iweka, K. E. Madu, and E. D. Edafiadhe, "Solar Radiation Forecasting Models and Their Thermodynamic Analysis in Asaba: Least Square Regression and Machine Learning Approach," *Journal of Energy Research and Reviews*, 2024, doi: 10.9734/jenrr/2024/v16i2333.

[26] X. Zhang and X. Xiang, "A Study on the Transaction Price of Second-Hand Sailboats Based on the Random Forest Regression Model," *Advances in Computer Signals and Systems*, 2023, doi: 10.23977/acss.2023.071001.