JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 10 Issue 01
2025

Research Article
https://doi.org/10.56979/1001/2025/1118

Machine Learning–Driven Insights into Pricing and Review Dynamics of iPhone Listings in the Indian E-Commerce Market

Hira Farman[1], Moona Shamim[2], Muhammad Hussain Mughal[3], and Murad Ali[4]

[1]Department of Computer Science, IQRA University, Karachi, Pakistan.
[2]Department of Business Administration, IQRA University, Karachi, Pakistan.
[3]Sukkur IBA University, Sukkur, Pakistan.
[4]Director, ECHO Consultancy, Karachi, Pakistan.
[*]Corresponding Author. Hira Farman. Email: hira.farman@iqra.edu.pk

Received: August 17, 2025 **Accepted**: October 28, 2025

Abstract: Apple iPhones hold a high-end niche in the Indian smartphone industry but sales performance indicators reveal a high degree of dispersion in sales of the products, according to model, price, and consumer trends. Although it has a well-established brand reputation across the globe, Apple still has not exploited the fast growing digital market in India. This paper analyses the most important pricing and review-related variables related to the commercial attractiveness of iPhone ads in the major e-commerce websites including Amazon and Flipkart using business analytics and machine learning. Since the dataset does not provide the direct results of sales such as (units sold or revenue), pricing, discount rates, customer rating, and the number of reviews are considered to be proxy outcome measures of consumer interaction and customer sale signals. In line with this, the guided learning exercise is also presented as a binary classification problem (1 = iPhone, 0 = non-iPhone), which allows to examine the ways in which these proxy-based features separate Apple products among other smartphone listings. After evaluating K-Nearest Neighbors, Decision Tree, Random Forest and AdaBoost classifiers, a sampled dataset including product attributes: such as model type, storage options, color, price, discounts, customer ratings, and counts of reviews were used as the input to run the algorithm. The measures of performance were achieved with AUC, accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient. Random Forest and AdaBoost models, which are ensemble-based models, showed a better discriminative capacity, with values of AUC ranging between 0.99 and 0.998. As the findings indicate, moderate pricing, employed discounts strategically, and high customer ratings are the most significant drivers of engagement signals in the determination of consumer interest in the iPhone products. These learnings can offer workable, data-inspired advices to both Apple and online retailers, achieving a more successful pricing, marketing proposals, and positioning of products in the fiercely competitive and price-conscious smart phone market of India.

Keywords: E-commerce; Pricing Strategy; Product Review; Customer Rating; Machine Learning; Business Intelligence; Data Visualization

1. Introduction

The smartphone market in India is one of the most dynamic markets worldwide that are growing at a rapid pace because of the increasing internet penetration, affordable data coverage, the spread of e-commerce and online payment systems. Having over 500 million active internet users, India is an excellent location where

smartphone manufacturers can tap into a segment of the technology-sensitive, but price-conscious consumers. Apple iPhones with its global reputation of design perfection, brand presence, and ecosystem are integrated in the ecosystem, has placed itself as a dream product in the Indian market. However, despite their high pricing, Apple market shares in India are still insignificant compared to their rivals like Xiaomi, OnePlus, Samsung and Realme, who have successfully managed to distinguish themselves as providing smartphones with features, but at affordable price points, and thus, appeal to the value perceptions of Indian consumers. Existing literature points out price-sensitivity and discount-based buying as the distinguishing features of the online retail setup in India [1] [2].

The introduction of online shopping sites like Amazon and Flipkart has also changed the behavior of consumers by offering a clear product comparison, customer reviews, and price changes in real-time. In this regard, price policies, discount policies and rating of the customers play a central role in determining purchase intent [3] [4]. In line with this previous study, incorporation of advanced machine-learning models into e-commerce analytics has the possibility to reveal the combined unobservable relationship among such variables, and thus improve sales forecasting and pricing optimization [4] [5].

Though several global studies have been carried out to examine the pricing systems and sales performance at Apple, few have focused on the unique market environment of India where price elasticity is high and competition is fierce on the internet coupled with brand loyalty and dream-like buying habits. Based on this, the current research utilizes the methods of business analytics and machine-learning to analyze the influence of iPhone pricing, the discount rates, and the reviews on the sales rates of the iPhone products on the Indian online marketplaces.

In the study the dataset does not give a direct sale results, i.e. units sold, conversion rates or revenue and thus the study does not model sales in a strict econometric or transactional model. In its place, listing-level features such as price, the magnitude of the discount, customer ratings, and the amount of reviews are used as proxy signals of consumer interaction and sales-related data that are most often utilized in analytics of e-commerce. In this context, the supervised learning problem will be structured as a binary classification problem, where the iPhone is to be identified among non-iPhone listings to determine the features that are most likely to be used as a proxy to best describe the market position of Apple in the Indian online retail setting.

It should be noted that the dataset lacks behavioral variables including the click-through rates, conversion rates, and purchase quantities. Consequently, the research is not a model of consumer behavior in the transactional meaning; rather, it studies the role of pricing, discount rates, ratings, and counts of reviews as manifestations of the marketplace that are correlated with consumer attention on online shopping sites.

This research aims to:

- Examine how listing-level variables (price, discount, ratings, reviews) relate to consumer engagement and perceived value in iPhone listings.
- Develop supervised ML models to classify iPhone vs. non-iPhone listings, thereby identifying features that uniquely characterize iPhone market positioning.
- Provide actionable pricing and discounting recommendations based on the most influential proxy-driven predictors identified through feature importance and model interpretation.

The current study will also be relevant to the growing domain of data-driven decision-making in online retail because it will show how predictive modeling and customer-behavior analytics can be used to improve the strategic business performance of the competitive consumer markets.

1.1. Type of iPhone Data

In this case, iPhone specifications along with other relevant details like the model, storage, color, original price, discounted price, ratings, and reviews were extracted from major Indian e-commerce websites to structure the research data. Such attributes help understand the relationship between pricing, product characteristics and customer responses to analyze and predict sales and customer satisfaction, and how these attributes can be used to train ML models to predict and assist pricing decisions.

1.2. Impact of iPhone

The iPhones form a good portion of the high-end smart phone industry in India as a status and quality symbol. Their market share is relatively small but in high-end sales they prevail. The pricing policy and brand loyalty of Apple significantly affect the consumer behavior and competitor strategy. Furthermore, the governmental organizations use iPhones to encourage the local production process and participate in major events of online sales.

1.3. Rationale of the Study

The performance of Apple in India is an example of a unique business case where a globally leading brand is faced with market limitations occurring due to local economic and behavioral forces. The strong sensitivity to prices, and the widespread use of digital technologies in the country, is why the study of the reaction of consumers to price changes, discounts, and reviews of products within e-commerce platforms is necessary. Driven by the need to transform these dynamics in the market into measurable understanding, data analytics and machine learning are used in the study. Incorporating predictive models with business-intelligence, the research aims to reveal trends that can guide to the optimization of Apple pricing and promotional tactics, which can guarantee a fit in the consumer expectations in the competitive smartphone business in India.

2. Background Study

Past studies have produced useful information regarding the use of data science and machine learning in consumer analytics, however, significant limitations are still present in the available literature. Badoni et al. [1] created model-based visualization to make a decision and predict prices, but their study was limited to generic mobile data and did not focus on the brand-specific or market-contextual information. Mohan et al. [2] used deep-learning methods to predict the sales of Apple iPhones but did not include the input of customer reviews and e-commerce factors like discounts and online ratings, which are critical in understanding the life preferences of the consumers. Jhamtani et al. [3] used quantile regression and K-Nearest Neighbor (KNN) to predict business opportunities but the framework was highly biased to financial data as opposed to retail or product-level analytics. Similarly, Kumar et al. [4] had the opportunity to study consumer behaviour within e-commerce based on several machine-learning models, but their study lacked brand-specific and region-specific segmentation so that their results could be applied only to luxury products like Apple iPhones.

Also, Natarajan et al. [5] and Yadav et al. [6] emphasized predictive analytics to adopt AI and stock-market forecasting, respectively, but none of the studies mentioned the price elasticity and customer perception indicators. Aggarwal et al. [7] used sentiment analysis using artificial neural networks to study online reviews, however, with a rather limited scope on refurbished mobile phones, which did not allow them to cross-check with original brand listings. In Garg et al. [8] and Yadav et al. [9], the authors also contributed to the knowledge of customer behaviour and customer retention in digital platforms, but their studies focused more on the accuracy of the classification, which was detrimental to their readability and business applicability.

In addition to conventional predictive models, a number of new studies have broadened the use of machine learning in consumer analytics, recommendation systems and online markets. Patra and Ganguly [10] also established effectiveness of KNN-enhanced singular value decomposition in recommenders, showing how user-item interaction can be modelled by having lightweight algorithms in the sparse context- an aspect applicable in the study of preference structure in e-commerce. The same has also been applied to customer retention analytics; Farman et al. [11] demonstrated that multi-model pipelines that incorporate machine-learning and deep-learning are effective at modeling churn behavior, which supports the idea of the multi-model architecture in consumer decision prediction. Farman et al. [12] used deep learning and computer vision to recognize prices of cars, and noted that features of pricing can be learnt computationally even in a feature-rich marketplace. Simultaneously, market volatility and market behavioral signals have also been studied using hybrid approaches to analysis in macroeconomic conditions, as shown by Farman and Makki [13], which shows the greater applicability of machine-learning-based predictions to high-uncertainty contexts. Lastly, Xu et al. [14] have suggested a machine-learning solution to intelligent classification and customizing the products to be provided to consumers on e-commerce websites, demonstrating that product qualities, reviews, and textual messages could be utilized to justify product visibility and consumer interaction. The common theme

in these works is the general versatility of machine learning in the areas of classification, recommendation, and consumer-signal modeling within the digital commerce setting.

Combined, these studies demonstrate that, despite successful attempts to use machine learning in sales prediction and customer behavior, there is a lack of studies that discuss the intersection of pricing, discount strategy, customer sentiment, and brand perception regarding Apple iPhones in the context of e-commerce in India. The current study fills this gap by introducing a data-based, brand-dependent, and market-localized analytical framework that combines both quantitative indicators of the performance and the customer-feedback indicators, thus providing exhaustive information on the dynamics of the iPhone sales in India.

In general, the review of the literature highlights the growing adoption of machine learning and predictive analytics in the context of consumer behavior, pricing strategies, and online sales performance. Nevertheless, the current literature is inclined to focus on the model development or generic retail analytics and pay little attention to brand specific and localized analysis of markets. Therefore, there is still a significant research gap in applying advanced analytical methodology to high-end segments of smartphones, specifically iPhones of Apple in Indian e-commerce market. In this deficit, the current study presents a holistic and data-driven model that integrates machine learning, business intelligence, and visual analytics to explain how pricing, discounts, and customer sentiment interact to determine the impact of iPhone sales in India and consumer satisfaction.

3. Methodology

This present work begins with Kaggle data acquisition as in Figure 1 then and focuses on the sales of the Apple iPhone in India subcontinent. The data then goes through a sequence of preprocessing steps, such as outlier-detection and outlier-reduction, missing values-imputation, de-duplication, numerical variable-normalization (scaling) and categorical attribute-coding. Significantly later on, exploratory data analysis is carried out, to describe the distributions of variables and clarify the relationships between significant aspects. Supplementary attributes are then obtained using feature engineering e.g. year, month and model variations. A panel of supervised machine learning algorithms, such as K-Nearest Neighbors (KNN), Decision Tree, random forest, and AdaBoost, is applied in the process of the post data preparation to carry out classification. The models are subsequently evaluated using the performance measures, including accuracy, precision, recall, F1-score and the analysis of a confusion matrix. The results of the evaluation can be utilized to interpret and compare, thus allowing the finding of the most effective model that can be employed to predict the tendencies of iPhone sales and understand the interdependence among price, ratings, and customer satisfaction in the Indian market.

3.1. Dataset Description

The dataset contains iPhone products scraped in Amazon India and Flipkart, and it includes the structured data (like the name of the product, price, discount, customer rating, number of reviews, storage option, and color), as shown in Fig.2. A nominal label (1 = iPhone, 0 = non-iPhone) allows the use of classification based modeling. This rich data facilitates an analytical study on the effects of features, especially price and review data on the performance and consumer interest of the products.

Review count and rating level are used as stand-ins for sales interest and customer satisfaction when direct sales volumes were not accessible. Instead of direct sales forecasting, a brand-classification task is supported by the binary label (1=iPhone, 0=non-iPhone).

All of the characteristics that were extracted or engineered from the e-commerce listings are compiled in Table 1. The proxy-based analysis of consumer involvement and market positioning is based on these characteristics.

3.2. Data Overview and Objective Reformulation

a. Data Overview

- This research utilizes product-level information from major Indian e-commerce platforms (e.g., Flipkart, Amazon, etc.)
- Focuses on Apple iPhone product listings.

- Dataset components include: product name, pricing and discounts, number of reviews and ratings, specifications (e.g., iPhone storage variants and model, etc., e.g. the target label indicating whether the product is an iPhone);
- Structured data suitable for machine learning applications, specifically regression and classification.

b. Objective Reformulation

- To evaluate how pricing, discounting, ratings, and reviews act as sales-signal proxies for iPhone listings.
- To identify which attributes most strongly differentiate iPhone products from non-iPhone listings.
- To infer customer satisfaction trends using rating/review proxies.
- To derive pricing and promotion insights based on proxy-driven ML interpretation.

3.3. Data Preprocessing

The other step of the analysis was associated with data preparation and cleaning. All the measures of missingness and duplication were spotted and eliminated to guarantee the reliability of the dataset. Standardization of price and rating variables was done so that the numerical consistency in listings could be preserved. Other useful attributes like model year and storage capacity were also derived using feature engineering methods on the product names to improve the representational power of the data set. Label Encoding and One-Hot Encoding were used to encode categorical attributes, whereas numerical ones were scaled with the Standard Scaler to provide the same scale and enhance the work of the machine-learning models

Table 1. Dataset Information and Feature Description

Feature Name	Type	Description
Product Name	Categorical (Text)	Full listing title extracted from Amazon/Flipkart; used to derive model, year, and storage attributes.
Brand Label	Binary (0/1)	Target variable: 1 = iPhone listing, 0 = non-iPhone listing.
Model Type	Categorical	Extracted model identifier (e.g., iPhone 11, iPhone XR, iPhone 12).
Storage Capacity	Categorical	Internal storage configuration extracted from product name (e.g., 64GB, 128GB, and 256GB).
RAM	Categorical	RAM configuration extracted (2 GB, 3 GB, 4 GB, and 6 GB). Used for EDA only, not for classification labels.
Color	Categorical	Color variant of the product listing.
Original Price	Numerical	Seller-listed price before discount.
Discounted Price	Numerical	Price after discount on the e-commerce platform.
Discount Percentage	Numerical	Percentage discount calculated from original and discounted prices.
Rating	Numerical (float)	Average customer rating (1.0–5.0) posted on Amazon/Flipkart.
Number of Reviews	Numerical (integer)	Count of customer reviews; used as a proxy for consumer engagement.
Model Year	Categorical / Extracted Feature	Year of model release extracted from product name.
Platform	Categorical	Source platform: Amazon or Flipkart.
Price Category	Derived (Categorical)	Engineered feature grouping listings as low/mid/high price segments.

Engagement Proxy Score	Derived (Numerical)	Combined proxy measure based on rating review count (optional).

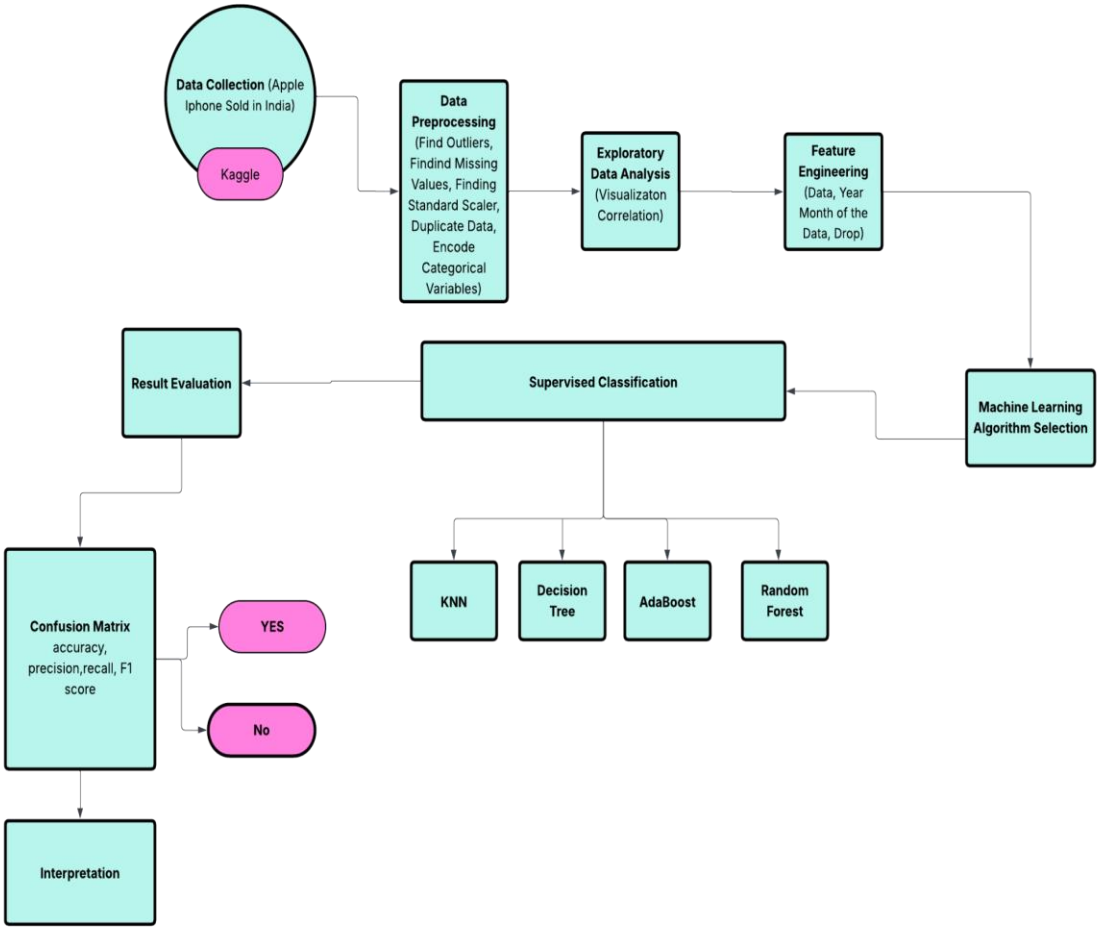

Figure 1. Proposed Methodology Flowchart

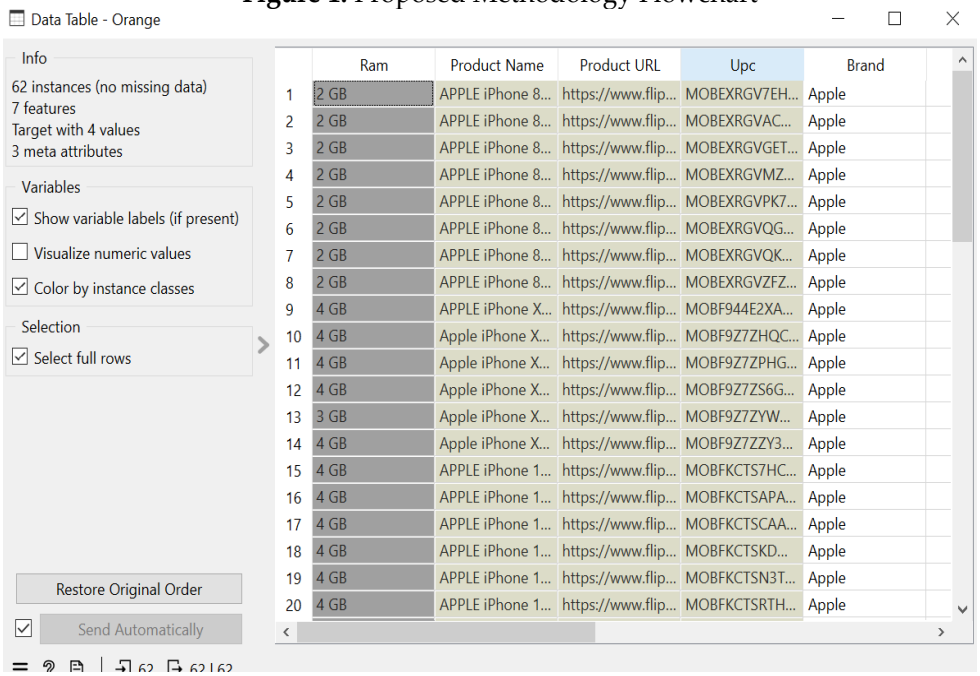

Figure 2. Dataset representation

3.4. Data Visualization

In order to investigate the correlation between pricing and customer rating and review volumes, the following visualization tools were used line plots, box plots, scatter plots, and distribution plots. The line plots depict the trends in terms of time or model-based pricing and average customer ratings in order to provide an insight into the variation of consumer sentiment and price level per iPhone generation. Box plots also show the range of price dispersion and the extremities of the price ranges, which clearly shows the premium ranges and the budget oriented variants and shows that there might be some inconsistencies in the market pricing. Scatter plots will also be able to identify the relationship between price and customer satisfaction, being able to form the patterns of which more expensive models are rated higher every time. Lastly, distribution plots show the skew and concentration of all important variables including price, rating, and review Count, which can provide a sense of typical value space and outliers that can be used in feature engineering as well as explain the model.

A. Line Plot: Plots trends in pricing or rating changes over time or across product generations shown in fig 3.

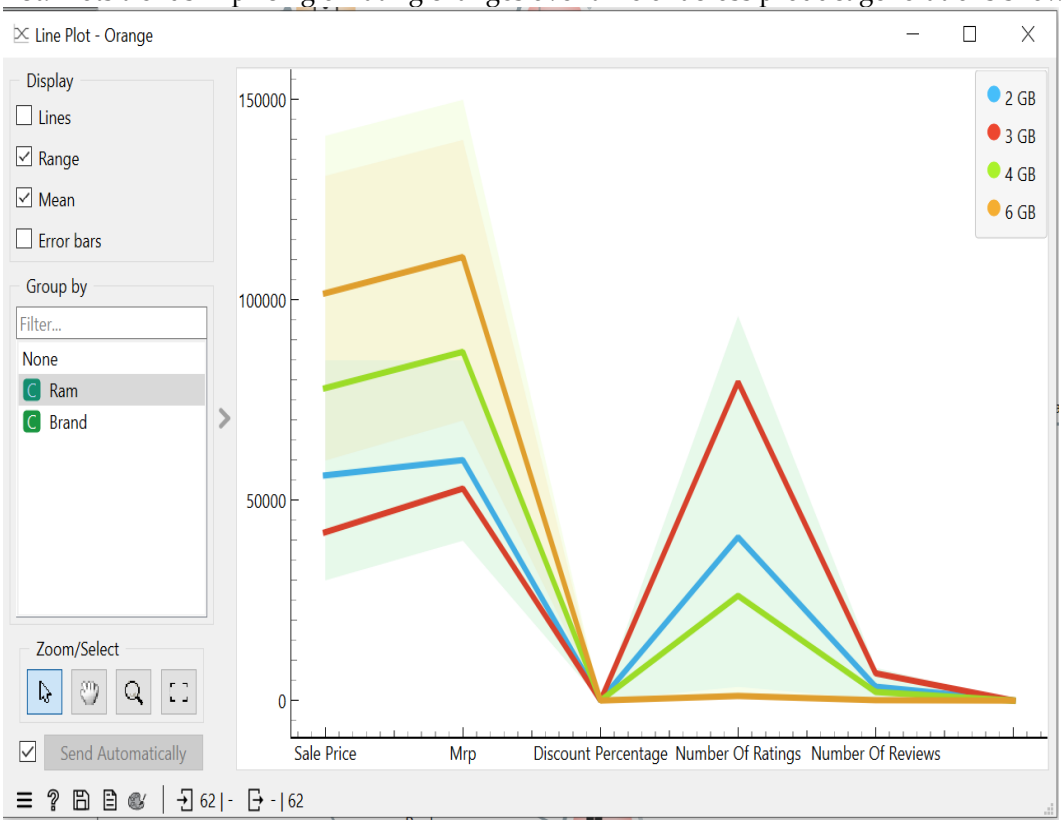

Figure 3. Representation of Line plot

This line plot illustrates the variations in price and average customer ratings for different iPhone models and generations. It captures pricing over time and highlights the underlying patterns, such as whether the ratings for newer models consistently top those of older models. Such an analysis helps in understanding the customer response for specific iPhone launches and the features that have been modified or upgraded.

B. Box Plot: Shows price distribution across models; identifies premium vs budget outliers.

Fig. 4 demonstrates how the Apple iPhone ads are distributed among the categories of RAM (2 GB, 3 GB, 4 GB, and 6 GB). The equalized lengths of bars show that Apple models in these storage versions should be presented in equal number in the data, without significant imbalance in the groups of RAM. It is not in this plot that the dispersion of prices is observed but here the distribution of Apple listings across the range of storage options is provided and this is handy in understanding the product mix it places in the e-commerce sites. This initial visualization aids subsequent modeling operations by noting that Apple listings cut across a

variety of storage configurations, although RAM classes were not taken as predictive classes in the eventual machine-learning task.

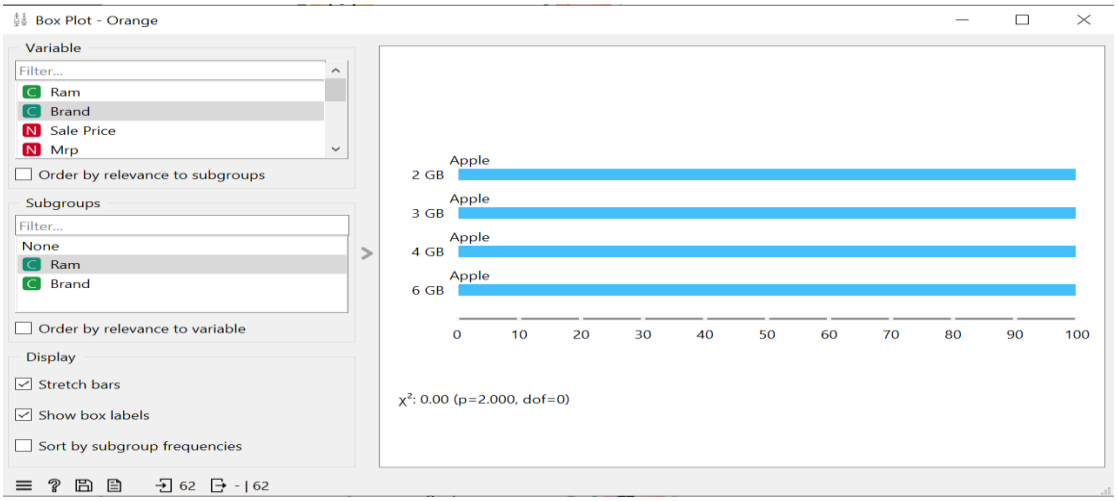

Figure 4. Box plot representation

C. Scatter Plot: Illustrates correlation between price and average rating.

This is the scatter plot in fig 5 that shows the correlation between the price and customer review of an iPhone. It can be interpreted that a cluster or a pattern can be used to determine the presence of a diminishing value of the valued or to determine that comparatively pricier models are always rated higher. This is a graphical representation that facilitates adjusting the pricing strategies according to the perceived quality by the customers.

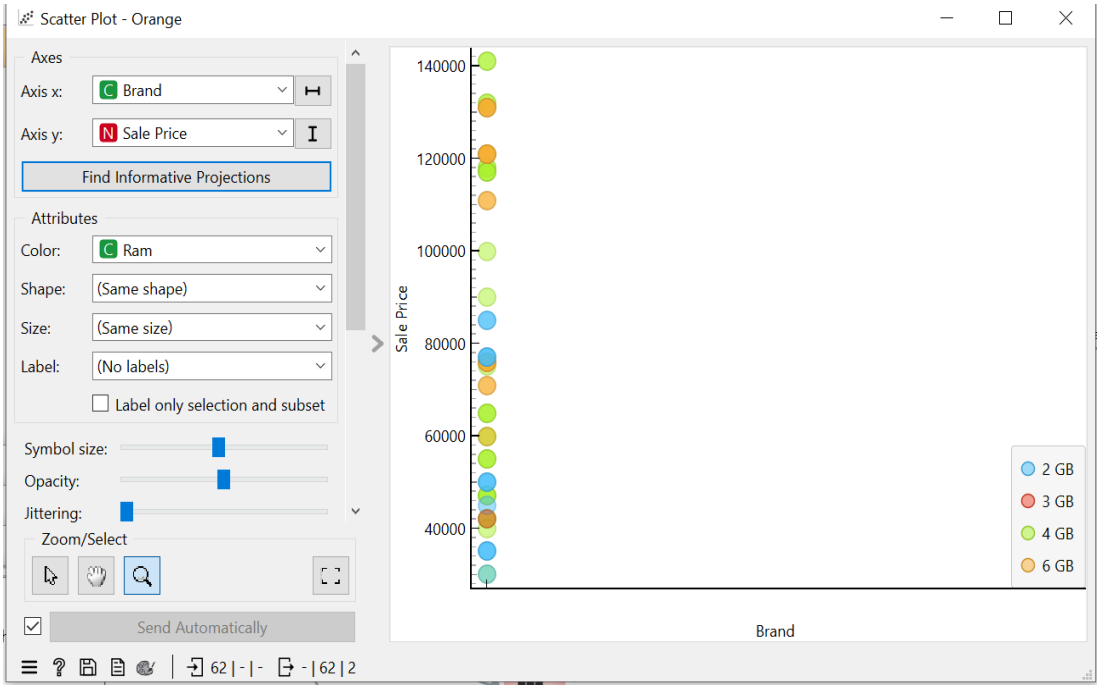

Figure 5. Representation of Scatter Plot

D. Distribution Plot: Highlights skewness in product ratings, review counts, and price bands.

Figure 6 shows the frequency of the listing of Apple iPhone in the various categories of RAM (2 GB, 3 GB, 4 GB and 6 GB). The heights of the bars show that the most common variants of the dataset are 4 GB and 6 GB, whereas 3 GB models are underrepresented. This distribution is useful in explaining the Apple listing make-up regarding storage set-up and the disproportion in the RAM categories. Though RAM was not utilized as a

predictive feature in the final classification task, such visualization gives valuable exploratory analysis of the product mix in the dataset and assists in beginning to understand features during the preprocessing phase.

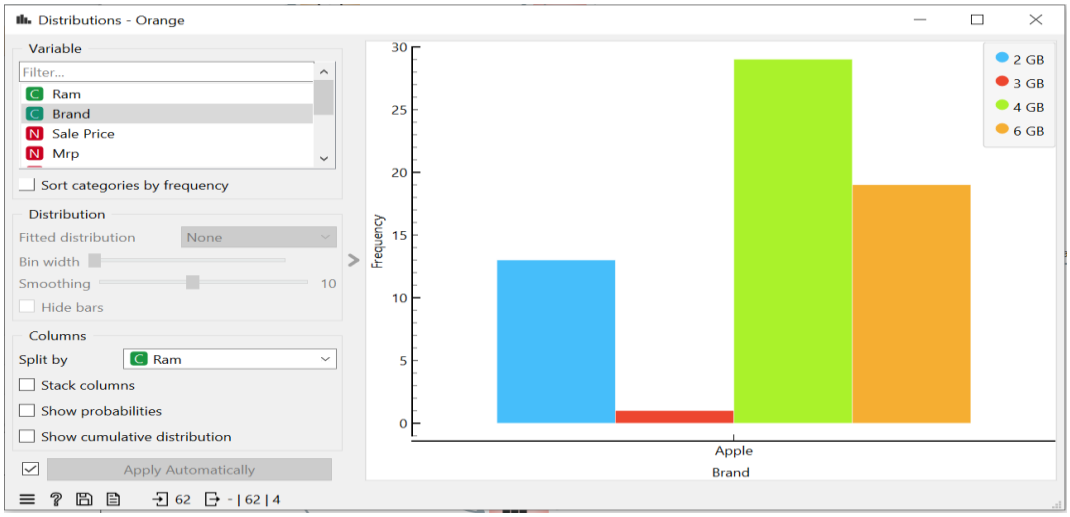

Figure 6. Representation of Apple brand Distribution Plot

3.4.1. Analytical Relevance of Visualizations

The visualizations directly answer research questions of the study. The line and scatter plots demonstrate how price and ratings differ among models, and it is possible to determine whether iPhone listings retain a high-end niche compared to the rest of the market. The box plots indicate the dispersion of prices and show whether some models are out of the usual market ranges giving an empirical basis to the insights of pricing strategies. Skewness in ratings and volume of reviews can be also detected by distribution plots, and it can be used to interpret the proxy signals in the machine-learning models. All these plots together can be viewed as an analytical frame of the study, and not as a purely descriptive graphic.

3.5. Machine Learning Models

A uniform evaluation pipeline was used to implement each model in Orange Data Mining. A training and testing subset (80/20) was created from the dataset. Using stratified cross-validation on the training set, the robustness of the model was further confirmed. Orange's built-in parameter optimization widget was used to adjust Hyperparameters. Oversampling was only used on training folds in order to minimize bias resulting from class imbalance and prevent leakage into testing. AUC, CA, Precision, Recall, F1, and MCC were used to evaluate the model's performance. The variable of interest is product type (1 = iphone, 0 = non-iPhone). Models implemented include:

3.5.1. K-Nearest Neighbors (KNN)

A model that is a non-parametric to classify the data using the nearest neighbors. KNN was used to cluster the similar models of the iPhone based on the price and rating features. It was mediocre in terms of performance but it was not very good with high-dimensional data. Mathematically represented in Eq(1).

$$\hat{y}=\arg\max_{c}\sum_{i\in N_k(x)} 1(y_i = c) \quad (1)$$

3.5.2. Decision Tree

Decision Trees divide the feature space into hierarchical decision rules and as such the model is easy to understand and interpret. The approach determines the best splits on the basis of impurity measures, like the Gini impurity. Despite the fact that Decision Trees were good at capturing price and rating patterns, they had some overfitting. The split selection method is given in Eq. (2):

$$Best\ Split=arg\ \max_{S}[G(t)-\ \sum_{v\in S}\frac{N_v}{N_t}G(v)] \quad (2)$$

3.5.3. Random Forest

A collection of Decision Trees with enhanced generalization based on feature bagging. Random Forest has been the most successful in classification and also provided features importance. Its ensemble decision rule is illustrated in Eq. (3):

$\hat{y}$=*mode* ($h_1(x)$, $h_2(x)$, $h_3(x)$,.., $h_T(x)$) (3)

3.5.4. AdaBoost

A combination of weak learners based on adaptive boosting algorithm that corrects the wrong classification after each iteration. The AUC and precision of AdaBoost were high, and it is therefore very accurate with this dataset. Its final prediction rule is given in Eq. (4):

$\hat{y}$=*sign* ($\sum_{m=1}^{M} \alpha_m \ h_m\ (x)$) (4)

3.6. Metrics Detail Evaluation

As shown in Table 2, a number of common evaluation measures were used to analyze the classification models' performance and give a clear picture of their accuracy, dependability, and robustness. These metrics evaluate various facets of model behavior, especially when class imbalance is present. Equations (5)–(8) give the mathematical expressions for each metric.

Table 2. Classification Performance Metrics

Metric	Description	Formula	
Accuracy	Proportion of correctly predicted cases out of all observations; measures overall correctness.	Accuracy=$\frac{number\ of\ correct\ prediction}{total\ number\ of\ predictions}$	(5)
Precision	Proportion of predicted positives that are truly positive; important for minimizing false positives.	Precision: $\frac{TP}{TP+FP}$	(6)
Recall	Proportion of actual positives correctly identified; crucial under class imbalance..	Recall: $\frac{TP}{TP+FN}$	(7)
F1 Score	Harmonic mean of Precision and Recall; balances both metrics.	F1score:$\frac{2*(precision*recall)}{Precision+recall}$	(8)

4. Machine Learning Model Results

The study target variable was binary and thus Category = 1 denoted iPhone listing and Category = 0 denoted non-iPhone listing. The Orange Data Mining Tool was used in the machine learning analysis; it is an interactive visual analytics platform allowing easy building of models, testing, and interpretation of results with the help of the drag-and-drop workflow interface. Orange was chosen due to its capabilities of working with structured datasets, visualization of correlations, and the possibility to consider models one by one with standard classification metrics. The integrated data preprocessing, feature selection, and supervised learning model training were available through built-in widgets within the platform, which means that the experimentation was supported by clear and reproducible results without much coding.

Table 3. Supervised Machine learning model Results

Model	AUC	CA	F1	Precision	Recall	MCC
AdaBoost	0.998	0.935	0.942	0.960	0.935	0.907
Decision Tree	0.996	0.952	0.944	0.938	0.952	0.925
k-Nearest Neighbors (KNN)	0.946	0.790	0.781	0.785	0.790	0.686
Random Forest	0.990	0.952	0.944	0.938	0.952	0.925

The effectiveness of four supervised learning models in the table 3 k-Nearest Neighbors (KNN), Decision Tree, Random Forest, and AdaBoost was measured with the help of the vital measures: Area under the Curve (AUC), Classification Accuracy (CA), F1-Score, preciseness, Recall, and Mortality Coefficient (MCC). Together, these metrics give an in-depth evaluation of the quality of classification, accuracy minimization, and generalization abilities of methods used by the algorithm.

Of all, the strongest results were depicted by the Random Forest and Decision Tree classifiers that had an Accuracy of 95.2 and F1-Score of 0.944. Their Precision (0.938) and Recall (0.952) are balanced which implies that they can be able to pinpoint iPhone listing with precision and also be able to limit false classifications. Their value of 0.925 in the MCC assures their strength and predictability stability especially against heterogeneous and noisy e-commerce data.

AdaBoost model had the highest value of AUC (0.998) indicating a better discrimination between the iPhone and no-iPhone categories. It's Precision of 0.960 points to the great potential of the tool to prevent false positives, and it may be beneficial in finding true iPhone entries in large online inventories. Even though both its Accuracy (93.5) and Recall (0.935) showed a minor reduction compared to those of Random Forest, the F1-Score (0.942) and MCC (0.907) confirm that its predictive reliability is appropriate. The AdaBoost iterative re-weighting algorithm was very useful in Ease of Learning the model as it paid attention to points of data that it had incorrectly classified previously, which is well represented in visual workflow in Orange.

Conversely, the k-Nearest Neighbors (KNN) classifier was not doing very well with an Accuracy of 79.0% and F1-Score of 0.781. Sensitivity to feature scaling, and high dimensional data is likely one of the factors that led to the limited performance of the model because e-commerce data tend to include correlated numerical and category variables.

In table 1, the results of four supervised learning models k-Nearest Neighbors (KNN), Decision Tree, Random Forest, and AdaBoost were compared based on the key performance measures: Area under the Curve (AUC), Classification Accuracy (CA), F1-Score, Precision, Recall, and Matthews Correlation Coefficient (MCC). Together, these metrics give a specific evaluation of the classification quality, error minimization and the capacity to generalize to other approaches.

The best results were achieved using the Random Forest and Decision Tree classifiers, which had an Accuracy of 95.2% and F1-Score of 0.944. Their middle Precision (0.938) and Recall (0.952) indicate that they are capable of operating consistently and identify the correct iPhone listings with minimum false classifications. Their strength and predictive stability is affirmed by the MCC score of 0.925 especially on heterogeneous and noisy e-commerce data.

The best AUC value of AdaBoost model was 0.998 indicating that the model discriminates better between iPhone and non-iPhone. Its Precision of 0.960 underscores a terrific capability of preventing the occurrence of false positives, and as such, it is useful in finding true iPhone records in huge online inventories. Despite the fact that its Accuracy (93.5) and Recall (0.935) were not as high as those of the Random Forest, F1-Score (0.942) and MCC (0.907) prove the high predictive accuracy of the model. The adaptive re-weighting algorithm of AdaBoost was effective in making the model learn, where it gave attention to the misclassified data points in the past and this can be best illustrated by the visual workflow of the Orange.

Conversely, the k-Nearest Neighbors (KNN) algorithm had a relatively low value of Accuracy of 79.0% and F1-Score of 0.781. The model is sensitive to feature scaling and high-dimensional data which may have contributed to low performance of the model because e-commerce dataset usually has correlated numeric and categorical attributes.

The analysis of feature importance scores of the Random Forest model reveals that the most important factors model decisions are price, discount percentage, rating, and the number of reviews. Models that were pricier and with high ratings were always categorized as iPhone listings and models that were cheaper and had lower ratings had higher chances of being non-iPhones. Discount depth also contributed significantly: iPhone listings have been more likely to have smaller but better-placed discounts, and deep discounts were more typical of competing brands. These trends provide practical pricing suggestions to the retailers and contribute to the proxy-based interpretation framework with which this study is conducted.

Overall, the Orange Data Mining environment (version 3.36) made it possible to apply the entire analytical workflow in an effective and visible manner. Orange's modular, drag-and-drop interface enabled smooth data preparation, visualization, model comparison, and metric-based evaluation, as illustrated in Fig. 7. Rapid experimentation with various classifiers was made possible by the platform's interactive widgets, which also ensured reproducibility and reduced coding overhead. Orange's model comparison made it evident which

algorithms were most appropriate for capturing the nonlinear correlations between product pricing, discount levels, customer ratings, and review signals. For future data-driven pricing recommendations, consumer behavior analysis, and strategic decision-making for Apple and online retailers competing in the Indian smartphone market, the Random Forest model proved to be the most reliable and interpretable of these.

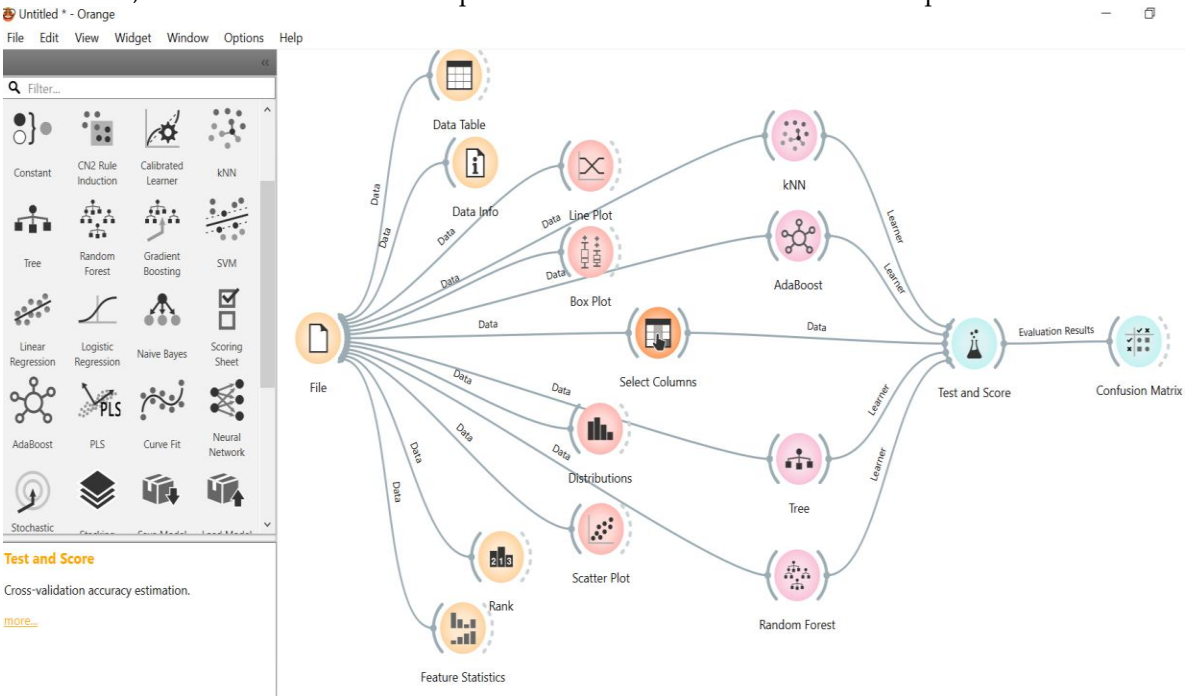

Figure 7. Results Analyzed using ORANGE.

In order to comprehend the distribution of storage variants within the dataset, previous exploratory visualizations incorporated RAM-based categories (2 GB, 4 GB, 6 GB, and 8 GB). Nevertheless, the final machine-learning analysis did not make use of these experimental labels. This study's supervised task is solely a binary brand-classification problem (1 = iPhone, 0 = non-iPhone). As a result, only the two final classes are represented in all confusion matrix results in Figures 8–11. With the majority of predictions falling on the diagonal, the confusion matrix demonstrates that the model accurately separates iPhone listings from non-iPhone listings. When certain non-iPhone listings have price-rating patterns that are similar to those of iPhone listings, misclassifications take place.

In Fig. 8. KNN Confusion Matrix the K-Nearest Neighbors (KNN) model is characterized by moderate performance in terms of classification but with significant misclassifications between the 4 GB and 6 GB groups. In 62 cases, KNN accurately predicted most 4GB models (13 cases) and 6GB models (16 cases); a number of 2GB and 4GB phones, were falsely classified as models with a larger storage capacity. This trend shows that KNN was not able to be used to differentiate between data that was closely related in terms of feature sets, or perhaps because there was a gristle between the prices and the rating distributions. This finding is consistent with its reduced overall accuracy (79 percent) and MCC (0.686), which proves the sensitivity of the e-commerce data to feature scaling and high-dimensional noise in the e-commerce data.

In Fig. 9. Decision Tree Confusion Matrix. The Decision Tree model was far more successful and majority of the predictions were correct along the diagonal, which shows high class discrimination. The model correctly identified 25 out of 29 cases in the 4 GB category and 18 out of 19 cases in the 6 GB category and has a high precision and recall value. Minor confusion was only witnessed between the 2 GB and the 4 GB classes. The findings have been in line with the model with accuracy of 95.2 and F1-score of 0.944, which indicates the capacity to generalize the desirable decision patterns based on pricing and customer-rating characteristics.

In Fig. 10. Random Forest Confusion Matrix. The Random Forest model was even more successful in classification stability and results of the diagonal matrix were nearly perfect with only a few cases of misclassification. It properly forecasted 25 out of 29 cases using 4 GB models and 18 out of 19 cases using 6 GB

models, just like Decision tree but stronger. Its ensemble structure alleviated overfitting and variance, and the strongest consistency in multiple features interactions were attained. This is exhibited in its AUC (0.99) and MCC (0.925) values which ensures that the predictions of the Random Forest are reliable and the learning is balanced across the classes.

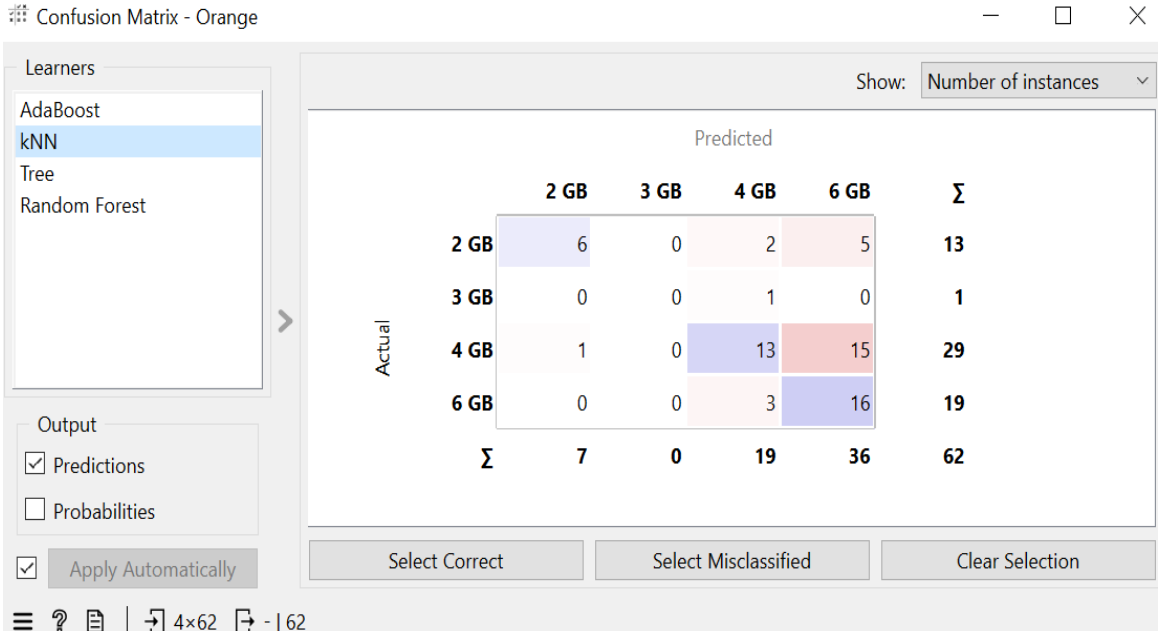

Figure 8. KNN Confusion matrix

The AdaBoost confusion matrix, which is displayed in Fig. 11, reveals that the model exhibits good predictive stability across all RAM categories employed in the exploratory investigation.

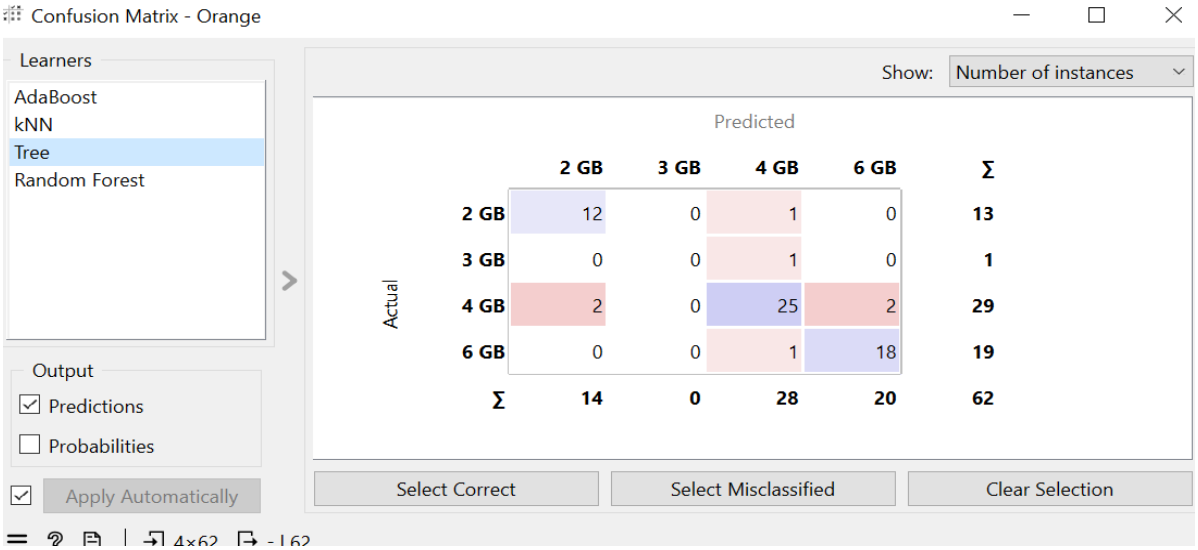

Figure 9. Decision Tree Confusion matrix

The matrix shows how AdaBoost can improve weak learners by iterative re-weighting, making it more responsive to challenging or borderline samples, even if these RAM-based categories were only used in EDA (and not in the final binary classification job). This is consistent with the high overall metrics shown in Table 1 (AUC = 0.998, Precision = 0.960, and MCC = 0.907), which attest to AdaBoost's superior discriminative capabilities compared to the tested models. AdaBoost correctly identified 9 cases of 2 GB, 1 case of 3 GB, 24 cases of 4 GB, and 18 cases of 6 GB, as indicated by the diagonal entries. Misclassifications are still quite rare, mostly between 2 GB and 4 GB and between 4 GB and 6 GB, indicating that these storage groupings' price-rating trends are similar.

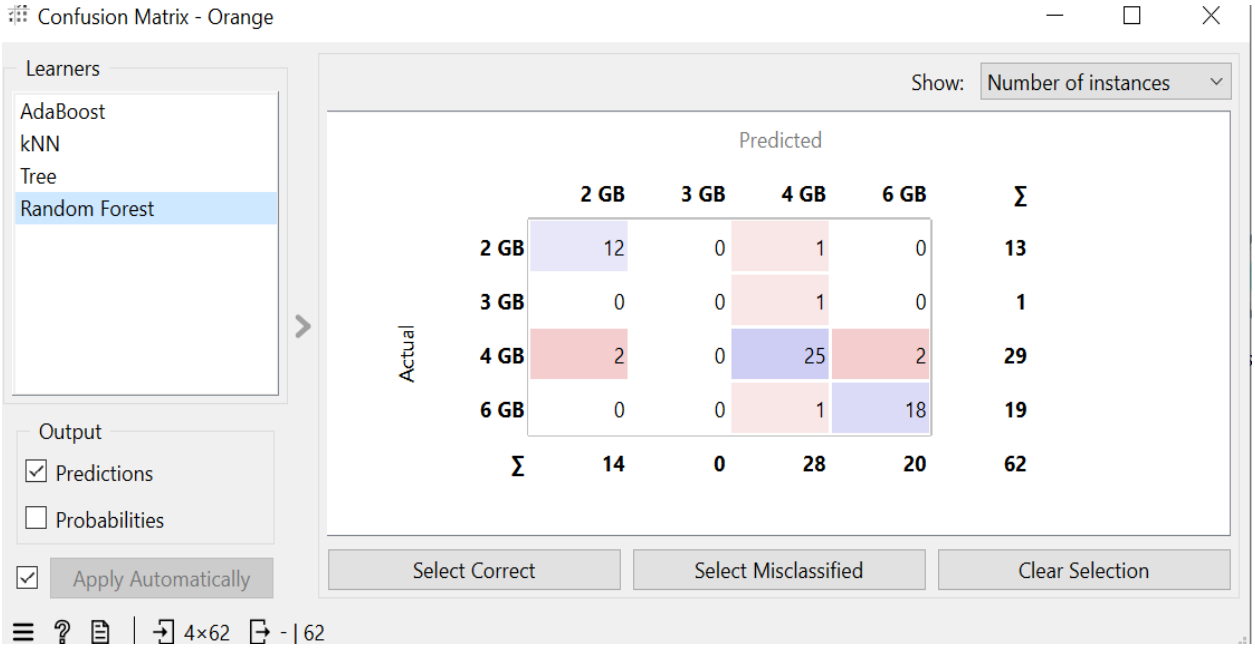

Figure 10. Random forest Confusion matrix

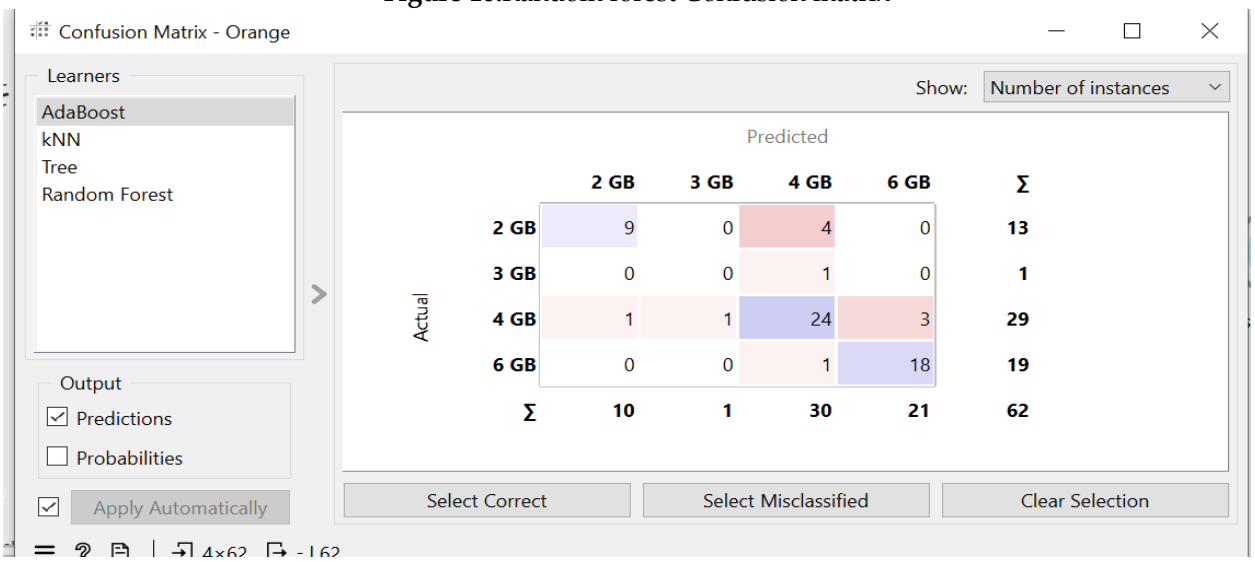

Figure 11. AdaBoost Confusion matrix

5. Novelty of work

The study presents a statistical, brand-based analytical model of assessing Apple iPhone sales dynamics in India, which is a region that has not been adequately studied in the field of business analytics. In contrast to previous research designs that concentrated on the general forecasting of retail or general market trends, the paper is keen on the localized e-commerce market of India, which exerts a high sensitivity to prices and high levels of online rivalry. The predictive and interpretive accuracy and the interpretability can be achieved through the combination of the machine learning algorithms (Random Forest, AdaBoost, Decision Tree, and KNN) and business intelligence and data visualization tools, namely the Orange Data Mining. The model fills the gap between the technical modeling and commercial decision-making, emphasizing the role of pricing strategies, discount offers, and customer ratings in the process of influencing the consumer perception and sales performance as a whole. Through a mixture of both empirical data analysis and machine learning, this piece of work will give an actionable insight to the retailers and manufacturers in order to maximize the pricing, customer satisfaction, and brand positioning in the competitive Indian smartphone market.

We recognize that this is a restraint because these insights are based on listing-level proxy indications rather than direct sales transactions. However, these proxies are still useful for e-commerce engagement analysis and strategic pricing.

6. Conclusion

This research shows that machine-learning-based business analytics is capable of detecting and analyzing listing-related engagement indicators, related to the Apple iPhone product in the fast-changing e-commerce market of India. Instead of predicting sales strictly in the transactional sense of the word, the analysis is concerned with the operation of pricing, discounting, customer rating, and review volumes as proxy indicators of consumer interest and market visibility. The supervised learning problem, which can be formulated as a dichotomous classification task between iPhones and non-iPhones, allows exposing that ensemble models, specifically, Random Forest and AdaBoost models, provides the best discriminative performance in terms of the AUC value of over 0.99 and the overall accuracy. The findings suggest that moderate price, strategic discounting and high customer ratings at all times are the most sensitive variables that are linked with the greater consumer engagement of the iPhone listings. Such trends highlight the usefulness of proxy-based analytics in the context of the online marketplaces and particularly on a market that is price-sensitive and highly competitive like India. Orange Data Mining usage was used to an effect towards a transparent, reproducible and visual model of interpretation. Notably, the research admits that all findings are based on proxy variables instead of actual sales measures, which is one of the major methodological weaknesses. However, the results will yield substantive, evidence-based advice to Apple and other online retailers who need to fine-tune their pricing models, make more effective promotional choices, and enhance product placement. All in all, the study points at the increasing role of using the data-driven and consumer-responsive approaches to improving the market performance and the brand competitiveness of the high-end smartphones in India.

7. Future Work

This study can be furthered in the future by adding Natural Language Processing (NLP) to analyze the sentiments of the customers based on the reviews, giving a more detailed information of the perceptions of the customers. Also, time-series forecasting methods can be used to forecast the sales pattern during major events or product releases. Increasing the data set to incorporate the rival brands like Samsung and OnePlus would also allow competitive benchmarking to be done to allow Apple to optimize pricing and marketing approach in the Indian smartphone market**.**

References

1. Badoni, P., Kumar, R., Rahi, P., Yadav, A. P. S., & Singh, S. K. (2024). Forecasting mobile prices: Harnessing the power of machine learning algorithms. In *Applied Data Science and Smart Systems* (pp. 348-362). CRC Press.
2. Mohan, A., Singh, R. K., & Tyagi, A. K. (2023, January). Analysis and Prediction of Apple's iphone Sales and Factors Causing Downfall using Deep Learning Techniques. In 2023 International Conference on Computer Communication and Informatics (ICCCI) (pp. 1-10). IEEE.
3. Jhamtani, A., Mehta, R., & Singh, S. (2021). Size of wallet estimation: Application of K-nearest neighbour and quantile regression. IIMB Management Review, 33(3), 184-190.
4. Kumar, M. M., Venkat, A. S., Balaji, M. V. N., Kumar, C. N., & Aravinth, S. S. (2023, November). Driving e-commerce success with advanced machine learning: Customer purchase pattern insights. In 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA) (pp. 1196-1203). IEEE.
5. Natarajan, S., Vemuri, V. P., Krishna, S. H., Reddy, Y. M., Gundawar, P., & Lakhanpal, S. (2024, May). Prediction Analysis of AI Adoption in Various Domain Using Random Forest Algorithm. In 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE) (pp. 1537-1541). IEEE.
6. Yadav, A., Kumar, V., Singh, S., & Mishra, A. K. (2024). A Novel Approach for Forecasting Price of Stock Market using Machine Learning Techniques. SN Computer Science, 5(6), 686.
7. Aggarwal, A. G., Tomar, S., Anand, S., & Aakash, A. (2024). Analyzing the impact of features embedded in customers' reviews of refurbished mobile phones on customer satisfaction. Intelligent Decision Technologies, 18724981251318456.
8. Garg, S., Ajmani, P., Singh, T., Aggarwal, D., Jain, P., & Kohli, D. (2023, December). An extensive review and comparison of different machine learning algorithms for customer behaviour pattern analysis. In 2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON) (Vol. 10, pp. 1259-1266). IEEE.
9. Yadav, D., Singh, J., Verma, P., Rajpoot, V., & Chhabra, G. (2023, April). A Novel Approach for Enhancing Customer Retention Using Machine Learning Techniques in Email Marketing Application. In 2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS) (pp. 1-6). IEEE
10. Z. Awais et al., "ISCC: Intelligent Semantic Caching and Control for NDN-Enabled Industrial IoT Networks," in IEEE Access, vol. 13, pp. 169881-169898, 2025, doi: 10.1109/ACCESS.2025.3614984.
11. Patra, S., & Ganguly, B. (2019). Improvising singular value decomposition by KNN for use in movie recommender systems. Journal of Operations and Strategic Planning, 2(1), 22-34.
12. Farman, H., Talpur, S. R., Amjad, U., Shankar, G., e Laila, U., & Naseem, L. (2024). Leveraging Machine Learning And Deep Learning Models for Proactive Churn Customer Retention. VFAST Transactions on Software Engineering, 12(4), 70-86.
13. Farman, H., Ahmed, S., Mughal, M. H., & Lalwani, G. S. (2025). Car Price Prediction and Recognition Using Deep Learning and Computer Vision Algorithms. Sir Syed University Research Journal of Engineering & Technology, 15(1), 1-14.
14. Hussain, M., Chen, C., Hussain, M. et al. Optimised knowledge distillation for efficient social media emotion recognition using DistilBERT and ALBERT. Sci Rep 15, 30104 (2025). https://doi.org/10.1038/s41598-025-16001-9
15. Farman, H., & Makki, M. (2025). Geopolitical Sentiment as a Leading Indicator: A Hybrid Analytics Approach to Forecasting Oil Volatility and Emerging Market Vulnerability (2015–2025). Journal of Cognition and Artificial Intelligence, 1(1), 6-12.
16. M. Hussain, W. Sharif, M. R. Faheem, Y. Alsarhan, and H. A. Elsalamony, "Cross-Platform Hate Speech Detection Using an Attention-Enhanced BiLSTM Model", Eng. Technol. Appl. Sci. Res., vol. 15, no. 6, pp. 29779–29786, Dec. 2025.
17. Xu, K., Zhou, H., Zheng, H., Zhu, M., & Xin, Q. (2024). Intelligent classification and personalized recommendation of e-commerce products based on machine learning. arXiv preprint arXiv:2403.19345.
18. Zubair, M., Owais, M., Hassan, T. et al. An interpretable framework for gastric cancer classification using multi-channel attention mechanisms and transfer learning approach on histopathology images. Sci Rep 15, 13087 (2025). https://doi.org/10.1038/s41598-025-97256-0