Residency project

Bereket Gebremariam, Aviinna Bhattarai and Roshan Acharya

[School of Computer Information Sciences](https://www.ucumberlands.edu/school-computer-information-sciences),  
University of the Cumberlands

MSCS-634

**Dr. Satish Penmatsa**

June 13, 2025

**Abstract**

This project presents a comprehensive data science analysis of the mobile phone market, leveraging machine learning models to extract business insights from a dataset of over 1,000 smartphone models. The study utilizes regression, classification, clustering, and association rule mining to identify key patterns in pricing, technical features, and market segments. Visualizations and metrics support the development of actionable strategies for product development, market segmentation, and competitive pricing. The ethical implications of data handling, model bias, and fair use are also discussed.

**Introduction**

Mobile phones are among the most dynamic and competitive products in the consumer electronics market. With rapidly evolving features and diverse pricing strategies, companies require data-driven tools to optimize product development and market positioning. This project applies modern machine learning techniques to analyze a dataset of 1,019 mobile devices across 15 variables. The objective is to uncover hidden insights that can support better business decisions in a competitive landscape.

**Methodology**

**Dataset Description**

The dataset includes mobile phone models with attributes such as RAM, internal storage, battery capacity, screen size, processor details, FM radio support, and price. Prices range from ₹99 to ₹489,990, reflecting a diverse mix of budget to flagship devices. Missing values were addressed through mode and median imputation, with categorical treatments for fields like FM radio availability.

**Data Cleaning and Feature Engineering**

Outliers were identified using the Interquartile Range (IQR) method, and duplicates were removed. Advanced parsing techniques (e.g., regex) were used to extract features such as battery size and RAM from unstructured text. Derived metrics like price-per-GB and performance scores were generated for additional insight.

**Exploratory Data Analysis**

Price distribution revealed a high positive skew with a mean of ₹35,007. Feature correlations indicated strong relationships between price and RAM (r = 0.72), battery capacity (r = 0.58), and storage (r = 0.65). Ratings were consistent across devices, with an average of 4.38 out of 5 and a narrow standard deviation, suggesting limited variation in user satisfaction.A screenshot of a graph

AI-generated content may be incorrect.  
Figure 1. Exploratory Data Analysis

A screenshot of a computer screen

AI-generated content may be incorrect.  
Figure 2. Correlation Matrix

**Machine Learning Analysis**

**Regression Modeling**

Three regression models were tested: Linear, Ridge, and Lasso. The Linear Regression model achieved the highest performance with an R² of 0.659 and an RMSE of ₹35,800, making it suitable for business pricing models. Key predictors included battery capacity, RAM, and performance score. These metrics directly translated to business insights (e.g., every 1,000mAh adds ₹1,270 to expected price)

A group of graphs with different colored bars

AI-generated content may be incorrect.  
Figure 3 Model Comparison

**Classification Modeling**

Using Random Forest, phones were classified into four price segments—Budget, Mid-range, Premium, and Flagship—with 65.46% accuracy. RAM and spec scores were the most important features. Confusion matrices revealed strong performance at the extremes (budget and flagship), with some overlap in mid-tier devices.

A group of graphs and charts

AI-generated content may be incorrect.  
Figure 4 Classification Model Analysis

**Clustering Analysis**

Hierarchical clustering (Ward’s method) identified three distinct customer segments:

1. **Budget-Friendly Mainstream** – Avg. price ₹22,000; 64% of dataset
2. **Premium Performance** – Avg. price ₹60,000; 30% of dataset
3. **Entry-Level Basic** – Avg. price ₹13,000; 6% of dataset

This segmentation supports tailored marketing and inventory strategies.

A group of graphs and diagrams

AI-generated content may be incorrect.  
Figure 5 Clustering Visualization and Analysis

**Association Rule Mining**

Over 2,200 high-confidence rules were discovered using association rule mining. For example, low RAM and low spec score strongly predicted budget pricing with an 81.7% confidence and 11.3 lift. Other rules revealed manufacturer-specific bundling strategies, supporting competitive intelligence and product planning.

**Business Recommendations**

* **Dynamic Pricing Engine**: Deploy the regression model for real-time pricing guidance with ±₹35,800 accuracy.
* **Segmented Marketing**: Tailor messages to three core clusters with distinct features and pricing.
* **Product Development**: Focus on battery and RAM as key pricing drivers (₹127/mAh and ₹89/GB respectively).
* **Feature Bundling**: Use high-confidence rules to design cost-efficient device configurations.

**Ethical and Bias Considerations**

The dataset contained no personally identifiable information, with all data aggregated from public specifications. Biases were identified in brand representation and feature availability and were mitigated through stratified sampling and fairness-aware validation. Fairness metrics confirmed consistent model performance across segments and brands, and transparency, accountability, and monitoring frameworks were proposed to support responsible AI deployment.

**Limitations and Future Work**

This analysis is limited by its static time scope and lack of market performance data such as sales or revenue. Future efforts could integrate temporal trends, customer reviews, and additional model types (e.g., deep learning). Applications like recommendation systems and real-time price optimization represent promising extensions of this work.

**Conclusion**

The project demonstrates the practical value of data science in the mobile phone industry. Machine learning enabled accurate price prediction, clear market segmentation, and the discovery of business-relevant patterns in device features. With ethical and fair practices in place, these insights offer a robust foundation for data-driven business strategies.