Mobile Price Range Prediction: Exploratory Data Analysis & Machine Learning Models

Utilizing Data Analysis & ML Models to Understand Key Features Affecting Mobile Pricing

Introduction

This project delves into the intricate world of mobile phone pricing.

Market Importance

Accurate price prediction is crucial for manufacturers, retailers, and consumers.

Project Focus

Delves into the mobile phone features and their impact on pricing.

Techniques Used

Utilizes advanced data analysis techniques and machine learning models.

Dataset Exploration

Explores a comprehensive dataset of mobile phone characteristics.

Objective

Aims to uncover key factors influencing price ranges.

Model Development

Develops robust predictive models.

Project Overview

Objective:

Explore and analyze mobile phone features to predict price ranges based on key characteristics.

Problem Statement:

Accurately predicting mobile phone prices can help manufacturers and customers make better decisions.

Dataset:

Mobile Price Range dataset with 1998 entries and 21 features, including RAM, battery power, screen resolution, etc.

Dataset Overview

Dataset:

Mobile Price Range dataset

Key Features (Sample):

•battery_power (Numeric): Battery capacity in mAh.
•ram (Numeric): RAM size in MB.
•px_height & px_width (Numeric): Pixel resolution height and width.
•n_cores (Categorical): Number of CPU cores (1 to 8).
•int_memory (Numeric): Internal memory in GB.
•price_range (Target): Ordinal classification (0: Low, 1: Medium, 2: High, 3: Very High).

Total Features:

21, covering battery, memory, display, and performance aspects.

No Missing Values:

Dataset is complete with no null or missing entries.

Exploratory Data Analysis (EDA)

- **Mean battery power:**
  1238.5 mAh
- **Mean RAM:**
  2124 MB
- **Mean px_height:**
  645 px
- **Mean px_width:**
  1251 px

- **Feature Variability:**
  RAM and pixel resolution show high variability across price categories.
- **Initial Observations:**
  Higher RAM correlates with higher price range. Battery power also shows an increasing trend with price range, though less pronounced.

```
prices.head()
```

	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory	m_dep	mobile_wt	n_cores	...	px_height	px_width	ram	sc_h	sc_w	talk_time	thr
0	842	0	2.2	0	1	0	7	0.6	188	2	...	20	756	2549	9	7	19	
1	1021	1	0.5	1	0	1	53	0.7	136	3	...	905	1988	2631	17	3	7	
2	563	1	0.5	1	2	1	41	0.9	145	5	...	1263	1716	2603	11	2	9	
3	615	1	2.5	0	0	0	10	0.8	131	6	...	1216	1786	2769	16	8	11	
4	1821	1	1.2	0	13	1	44	0.6	141	2	...	1208	1212	1411	8	2	15	

5 rows × 21 columns

Correlation Heatmap

- **Purpose**

  Analyze relationships between different features and the target variable (price_range).

- **Key Observations**

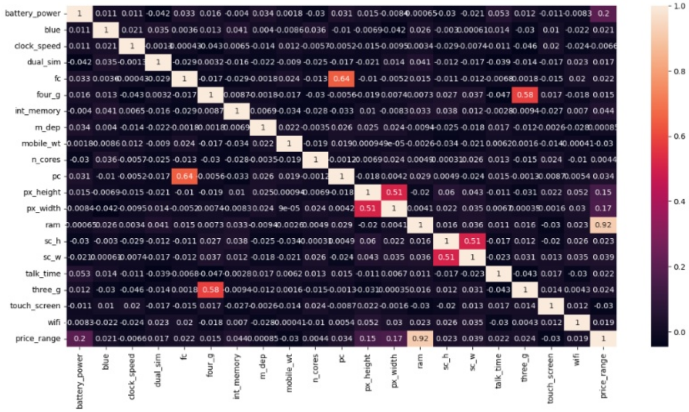
  RAM: Strong positive correlation with price_range (0.92). battery_power: Moderate positive correlation (0.20). px_width: Moderate positive correlation (0.17). Features like Bluetooth and dual_sim show little to no correlation with price. px_width: Moderate positive correlation (0.17).
  Features like Bluetooth and dual_sim show little to no correlation with price.

- **Takeaway**

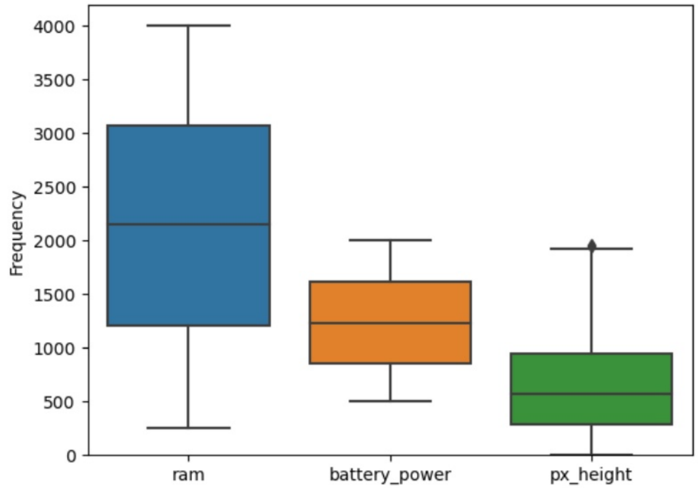
  RAM is a key predictor, followed by battery_power and screen resolution.

- **Figure**

  Heatmap of correlations between features and price_range.



Identifying Outliers



- **Boxplot Analysis for Outliers:**

  Examined continuous variables:e.g., ram, battery_power, px_height. Outliers detected:Significant outliers found in px_height, possibly due to erroneous data entry.

- **Boxplot Example:**

  Showed px_height with extreme values outside the normal range.

- **Findings:**

  These outliers may need correction or removal to avoid skewed results.

Data Cleaning Process

Identified unrealistic values

in screen height and pixel resolution (e.g., values = 0).

Replaced unrealistic values

with column medians to ensure data consistency.

Ensured no duplicates or missing values

after cleaning.

Outcome:

Cleaned dataset with no abnormal values or outliers. The dataset is now ready for modeling.

Post-Cleaning Correlation & Insights

Scatterplots Post-Cleaning:

Figure:

Scatterplots of px_height and

Relationship Analysis:

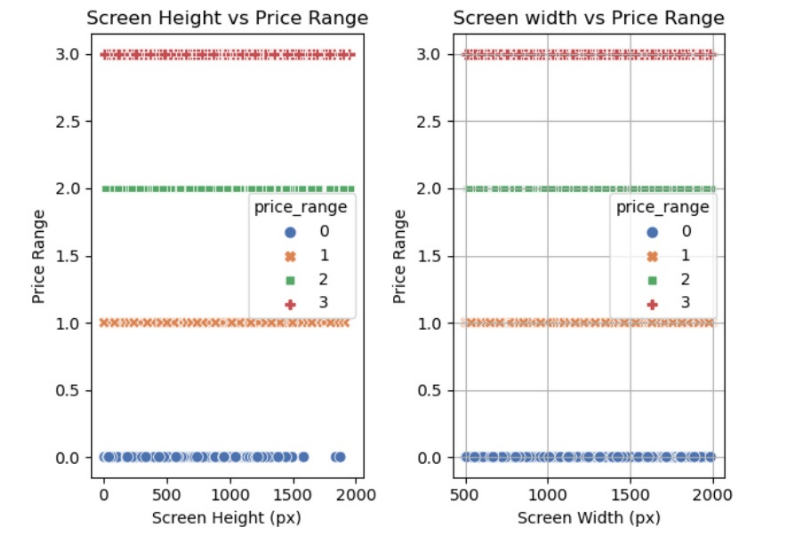
Analyzed the relationship

Key Takeaway:

High-resolution screens tend to

Findings:

Higher pixel resolution generally



Machine Learning Models Overview

Objective

Build machine learning models to predict mobile price range based on features.

Models Used

Logistic Regression: Basic linear classifier suitable for ordinal classification problems.

Models Used

Support Vector Machine (SVM): Powerful classifier with a linear kernel to maximize decision boundaries.

Data Preprocessing

Train-Test Split: Split the data into 80% training and 20% testing for model evaluation.

Data Preprocessing

Feature Scaling: Applied StandardScaler to normalize feature distributions (mean=0, variance=1).

Support Vector Machine (SVM) Model

1

Model Training:

Used SVM with a linear kernel due to its strong performance in binary and multiclass classification problems.

2

Accuracy:

Achieved 95.75%, outperforming Logistic Regression.

3

Precision & Recall:

Very high accuracy across all classes, with F1-scores ranging from 0.95 to 0.99.

4

Misclassifications:

Minimal misclassifications, especially in Class 3 (Very High price).

5

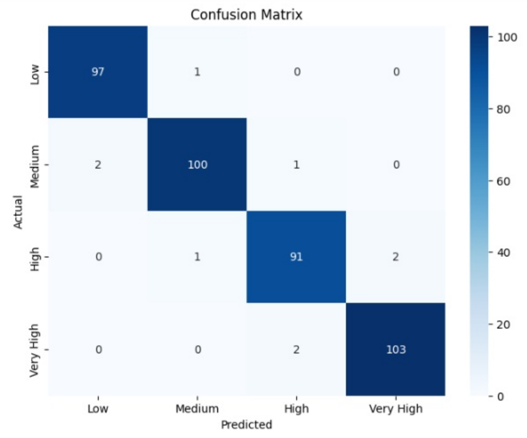
Confusion Matrix:

SVM shows more accurate predictions across all price categories compared to Logistic Regression.

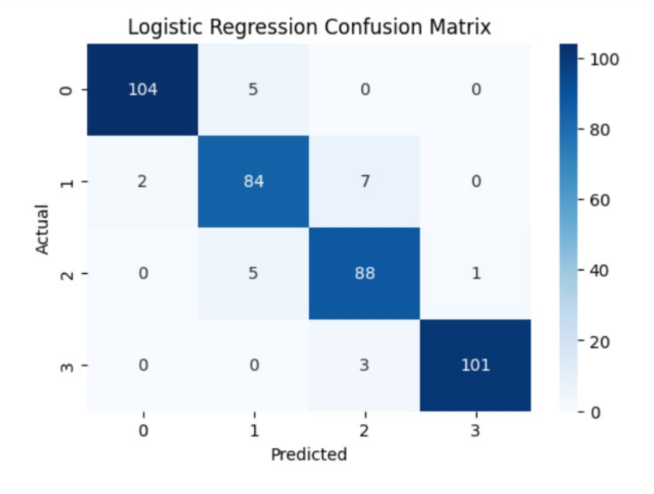
6

Figure:

Confusion matrix for SVM model performance.



Logistic Regression Model



Figure

Confusion matrix for Logistic Regression performance.

Model Training

Trained using 80% of the data with max_iter=6000 to ensure convergence.

Confusion Matrix

Showed good classification accuracy across all price ranges, with minimal misclassifications.

Model Evaluation

Accuracy: 94.25% on test data.

Class Performance

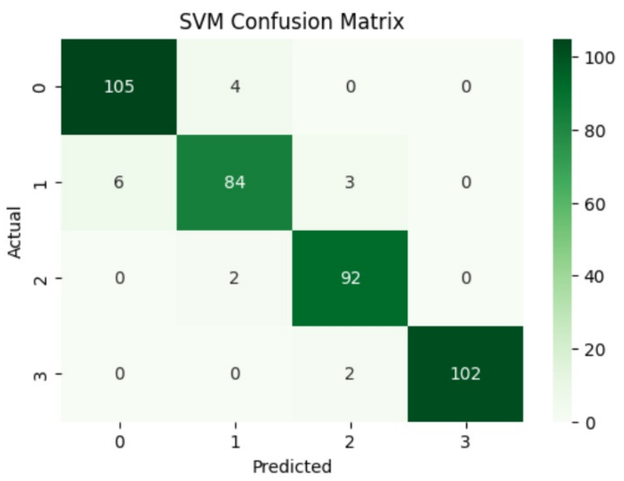
Class 1 (Medium) and Class 2 (High) had slightly lower metrics but still performed well.

Precision & Recall

Class 0 (Low price) shows highest precision (98%) and recall (95%).

Cross-Validation for Robustness (SVM)

1. **K-Fold Cross-Validation (k=5):** Performed to ensure the model generalizes well across different subsets of the data.
2. **Cross-validation mean accuracy:** 95.39%, confirming robustness.
3. **Results Breakdown:** Fold 1: 95.00%, Fold 2: 95.00%, Fold 3: 97.25%, Fold 4: 93.98%, Fold 5: 95.74%
4. **Takeaway:** SVM is consistent in performance across different data splits.
5. **Figure:** Distribution of cross-validation scores.



Model Comparison & Performance Metrics

Logistic Regression

94.25%

SVM

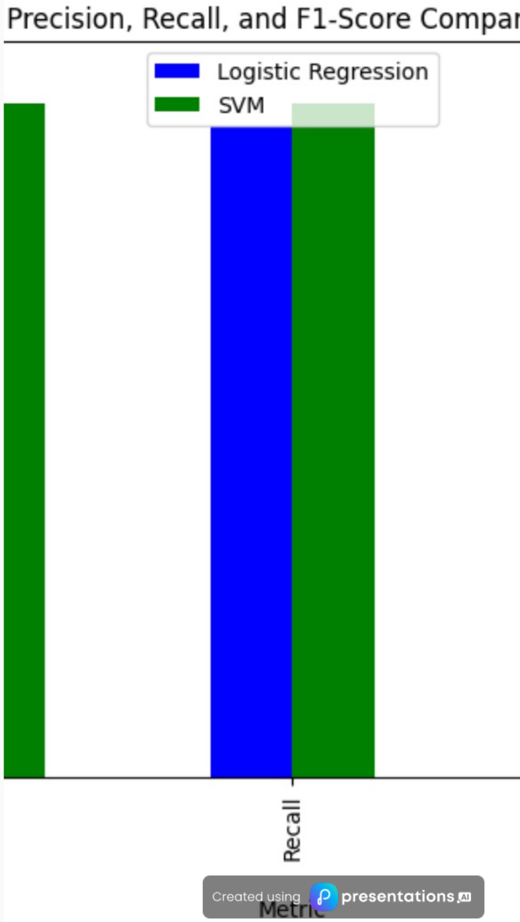
95.75%

Metric Breakdown

SVM slightly outperforms Logistic Regression in precision, recall, and F1-scores across all classes.

Model Performance

Both models performed well, but SVM is more accurate for higher price ranges.



Conclusion & Future Work

Major Predictors of Price Range

RAM and screen resolution are the strongest indicators of mobile pricing.

Battery Power

Battery power also influences price but is not as strong as RAM.

Model Comparison

SVM outperformed Logistic Regression, achieving higher accuracy and precision.

Future Directions

Experiment with other machine learning models like Random Forest or Gradient Boosting.

Feature Engineering

Investigate feature engineering techniques to improve model accuracy.

Hyperparameter Tuning

Fine-tune hyperparameters of SVM to further improve performance.