جامعة عفت

EFFAT UNIVERSITY

Predicting Mobile Phone Prices Using Machine Learning

CS 3072 - 1: Data Science Project Report

Leen Sharab - Lujain Almarri - Mawaddah Alagha

Instructor: **Dr Zain Balfagih**

December 15, 2024

Contents

0.1 Introduction

This project aims to analyze the relationship between mobile phone specifications, performance metrics, and their pricing in the competitive smartphone market. The dataset used in this study is called 'Mobile Dataset' [4] which comprises various features such as processor type, RAM, storage capacity, display size, camera quality, battery life, and brand. By employing machine learning techniques and data analysis, this project seeks to identify the key drivers of mobile phone prices and explore how these factors interact to influence pricing strategies.

Preliminary observations suggest that specifications like RAM, processor performance, and camera quality significantly affect pricing.

This report is structured as follows: first, the problem statement and background are presented to underline the significance of understanding mobile phone pricing. Then, the dataset and preprocessing methods are detailed. Following this, the analytical methods and machine learning models employed are discussed, along with key findings. Finally, the report concludes with insights derived from the analysis and suggestions for future research.

0.2 Background

Several studies have investigated the impact of specific smartphone features on pricing. For instance, researchers have analyzed how attributes such as camera resolution, battery life, and screen size affect price. However, there is a lack of comprehensive studies that consider multiple factors together or apply advanced techniques like machine learning for price prediction.

Key related works include:

- Regression models were used to study the relationship between smartphone specifications and price [5].
- Price prediction models employing decision trees and ensemble methods in e-commerce platforms[3].
- Analyses focusing on the importance of various features in consumer electronics, highlighting trends and consumer preferences[2].

This project builds on these existing efforts by integrating multiple features into a holistic analysis and leveraging machine learning to predict prices and uncover patterns in smartphone pricing strategies.

0.3 Research Question and Problem Statement

What are the most important factors influencing mobile phone prices, and how do these factors interact to predict price using machine learning?

In today's fast-growing smartphone market, understanding what makes mobile phones expensive is important for companies and customers. Phone prices depend on many things like features, performance, and brand, but how much each factor matters is not clear. This project looks at data about mobile phone specifications to find the key features that affect prices. Using data analysis and machine learning, we aim to discover patterns and connections, such as how RAM, cameras, or processors impact price. The findings will help companies design better products and guide customers in choosing the best phones for their needs[1]. [5]

0.4 Data

Unit of Observation: Each observation in the dataset represents a specific mobile phone model with its associated specifications and pricing details[4].

Outcome Variable: The target variable is **price_usd**, representing the price of the mobile phone in USD.

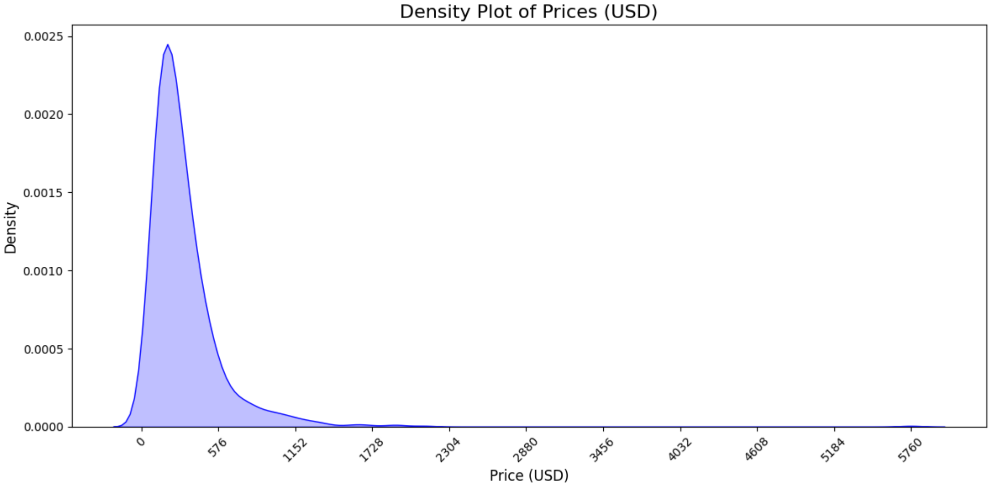

Figure 1: Density Plot of Mobile Prices (USD)

Figure 1 shows the distribution of mobile phone prices in the dataset. The data is right-skewed, indicating that most devices are priced at the lower end, with a sharp peak below $600. The long tail suggests the presence of premium-priced devices, extending up to $5760.

Predictor Variables: Predictor variables include a combination of raw and engineered features:

- **Raw Features:** rating, specs_score, ram_gb, inbuilt_storage_gb, battery_capacity_mah, fast_charging_w, front_camera_mp, display_size_inches, processor_speed_ghz, os, sim_type, and connectivity features (supports_3g, supports_4g, supports_5g).
- **Engineered Features:** price_per_ram, price_per_storage, performance_score, and ram_storage_interaction.

Table 1: Summary Statistics of Predictor Variables

Variable	**Count**	**Mean**	**Std. Dev.**	**Min**	**25%**	**Max**
Rating	785	4.37	0.24	3.45	4.15	4.75
Specs Score	785	80.42	6.64	60.00	76.00	88.00
RAM (GB)	785	7.43	2.76	2.00	6.00	12.00
Storage (GB)	785	163.73	84.85	16.00	128.00	256.00
Battery (mAh)	785	4903.88	524.47	1821.00	4800.00	5500.00
Fast Charging (W)	785	44.06	35.84	0.00	18.00	120.00
Processor Speed (GHz)	785	2.51	0.46	1.60	2.20	3.35
Price Per RAM	785	43.59	40.65	11.25	26.79	149.00
Price Per Storage	785	2.09	1.69	0.42	1.22	6.52
Performance Score	785	1599.79	875.43	195.20	960.00	3946.40
RAM-Storage Interaction	785	1377.22	1150.69	32.00	768.00	3072.00
Supports 3G	785	1.00	0.00	1.00	1.00	1.00
Supports 4G	785	1.00	0.00	1.00	1.00	1.00
Supports 5G	785	0.69	0.46	0.00	0.00	1.00

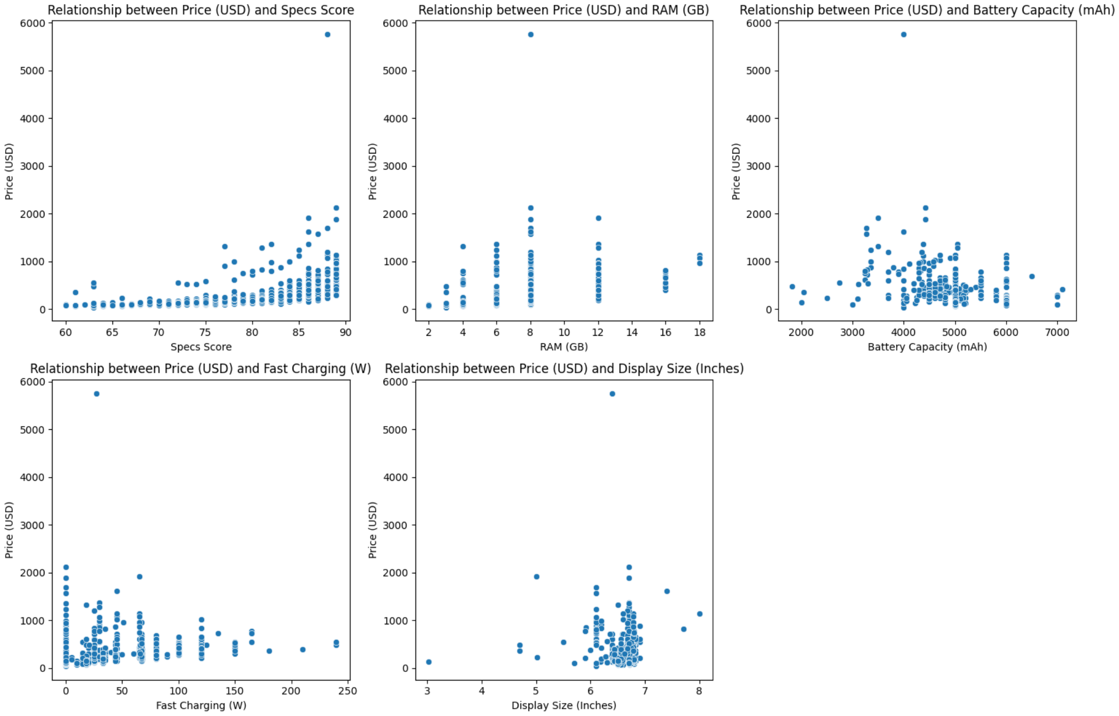

Figure 2: Density Plot of Mobile Prices (USD)

The scatter plots in figure 2 illustrate the relationships between mobile phone prices (in USD) and key features, including specs score, RAM, battery capacity, fast charging, and display size. A positive correlation is observed between price and specs score, as well as between price and RAM, indicating that higher specifications and larger RAM capacities often correspond to higher prices. In contrast, the relationship between price and battery capacity appears weak, with prices spread across various capacities. Similarly, fast charging capacity shows limited influence, as most high prices are concentrated at lower charging capacities, with a few outliers at higher levels. Display size shows a slight positive trend, with larger displays associated with marginally higher prices. Overall, these plots highlight the varying impacts of these features on pricing, with some factors showing stronger relationships than others.

Potential Issues with the Data:
The dataset has several issues, including missing values in key columns like ‘specs_score‘, ‘front_camera_mp‘, ‘battery_capacity_mah‘, and ‘core_type‘. Division by zero errors may occur during the creation of engineered features such as ‘price_per_ram‘ and ‘price_per_storage‘. Outliers in ‘price_usd‘ and other numerical columns could affect model performance. Additionally, non-numeric data in columns like connectivity, processor, storage, battery, display, and camera need to be split into separate columns and converted to numerical formats where applicable.

Solutions to the Issues:
Missing values in numerical columns like ‘specs_score‘, ‘front_camera_mp‘, and ‘battery_capacity_mah‘ were imputed with median values, while categorical columns like ‘core_type‘ were filled with default values such as ”Unknown.” Infinite values in engineered features were replaced with NaN, and rows with missing values were dropped. Exploratory data analysis was conducted to identify and interpret outliers rather than removing them to maintain data integrity. Non-numeric data was cleaned and transformed: connectivity was split into binary columns (‘supports_3g‘, ‘supports_4g‘, and ‘supports_5g‘), processor details were extracted into ‘core_count‘ and ‘clock_speed‘ (GHz), and storage was divided into ‘internal_storage_gb‘ and ‘expandable_storage_gb‘. Battery details were processed to include ‘battery_capacity_mah‘ and fast-charging features, display details were extracted into ‘screen_size_inches‘ and resolution, and camera data was split into ‘front_camera_mp‘, ‘rear_camera_mp‘, and camera count.

0.5 Analysis

0.5.1 Methods/Tools Explored

In this project, we employed Python libraries and tools to thoroughly analyze the ”Mobile Phone Pricing and Specification” dataset. Python was chosen for its robust capabilities in data analysis, feature engineering, and machine learning.

0.5.2 Key Python packages used included:

- **pandas and numpy:** For data manipulation, cleaning, and preprocessing.
- **matplotlib and seaborn:** For creating insightful visualizations to explore data relationships and distributions.
- **sklearn (scikit-learn):** For machine learning tasks including training models, evaluation, and feature importance analysis.

0.5.3 Analysis Approach:

The analysis included rigorous exploratory data analysis (EDA) to comprehend the dataset, address missing values, and explore potential correlations among variables. Feature engineering was extensively applied to create derived metrics that enhanced the predictive capability of the model. Given the characteristics of the dataset, the Random Forest Regressor was chosen as the primary predictive modeling technique due to its ability to handle non-linear relationships and capture complex data interactions effectively. The workflow involved multiple steps to ensure high-quality predictions and meaningful insights, supported by clear and interpretable visualizations of the data and model performance.

0.5.4 Detailed Analysis Outline

The analysis followed these key stages:

1. Data Preprocessing and Cleaning:

- Addressed missing values by imputing medians for numerical features and using placeholders for categorical features.
- Handled non-numeric data by converting categorical variables into binary or multi-level columns.
- Normalized numeric data where appropriate to ensure consistent scaling.

2. Exploratory Data Analysis (EDA):

- Visualized data distributions and relationships using histograms and scatter plots.
- Created a correlation heatmap to understand interactions between predictor variables.

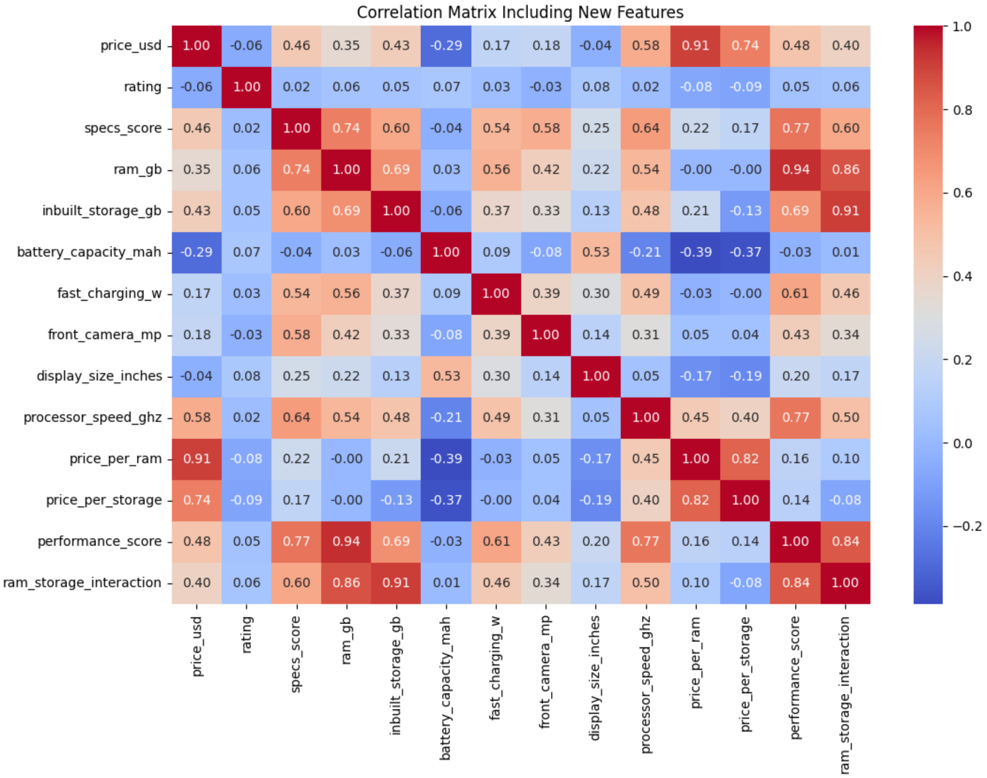

Figure 3: Relationships Between Predictors

Figure 3 highlights key relationships, such as the strong correlation between price_usd and engineered features like price_per_ram and performance_score, supporting the validity of the feature engineering process.

3. **Feature Engineering and Selection:**
   - Identified key predictor variables through correlation analysis and model insights.
   - Engineered new features, such as price_per_ram, performance_score, and ram_storage_interaction, to enhance model interpretability and performance.

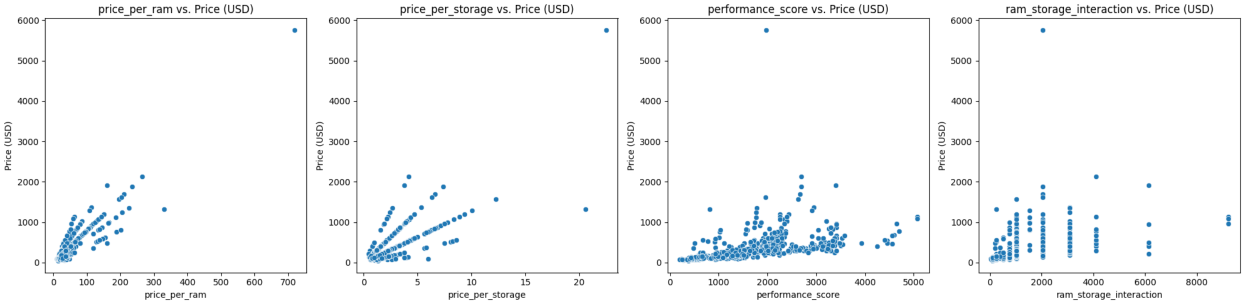

Figure 4: Relationship Between price in USD and new features

The scatter plots illustrate the relationship between key features and mobile phone prices. The plot for ‘price_per_ram‘ vs. Price (USD) displays a positive correlation, indicating that higher price-per-GB RAM is associated with higher overall phone prices. Similarly, the ‘price_per_storage‘ vs. Price (USD) plot shows a comparable trend, where price increases with the price-per-unit storage. The ‘performance_score‘ vs. Price (USD) plot reveals a strong positive relationship, suggesting that higher performance scores are closely tied to higher phone prices. However, the ‘ram_storage_interaction‘ vs. Price (USD) plot demonstrates limited correlation overall, with higher prices observed primarily at extreme values of the interaction term.

4. **Model Building:**
   - Trained a Random Forest Regressor for its robustness and ability to handle non-linear interactions.
   - Split data into training and test sets (80/20).
5. **Model Interpretation and Evaluation:**
   - Interpreted the model using feature importance scores to understand key drivers of price_usd.
   - Evaluated performance metrics, including $\mathbf{R^2}$ and Mean Squared Error (MSE), to quantify model accuracy.
   - Compared the performance of multiple models:
     - **Random Forest:** Achieved an R-squared of 0.97 and a Mean Squared Error (MSE) of 2039.74, demonstrating strong predictive power.
     - **K-Nearest Neighbors (KNN):** Resulted in an R-squared of 0.81 and a higher MSE of 12,544.53, indicating that while it captures trends, it is less precise than Random Forest.
     - **Gradient Boosting:** Outperformed the other models with an R-squared of 0.99 and a low MSE of 870.29, making it the most accurate model for the dataset.
   - Visualized results with predicted vs. actual prices and residual plots.
6. **Validation and Testing:**
   - Assessed model robustness on a separate test set to ensure generalizability.
   - Confirmed reliability using additional performance metrics, such as cross-validation $\mathbf{R^2}$ scores.

This structured methodology ensured a comprehensive understanding of the dataset, leading to robust predictive modeling and actionable insights.

0.6 Results

0.6.1 Summary of Results

The predictive analysis was conducted using a Random Forest Regressor as the primary model, leveraging its ability to handle complex, non-linear relationships within the dataset. The model was trained on a cleaned and processed dataset partitioned into training and testing sets.

1. Model Performance

- **Random Forest Model:** Achieved high accuracy with a Mean Absolute Error (MAE) of **16.93** and an R^2 score of **0.97** on the test set, indicating strong predictive performance.
- **K-Nearest Neighbors (KNN):** Demonstrated moderate accuracy with an R^2 score of **0.808314** and a Mean Squared Error (MSE) of **12544.526751**, showing some limitations in handling complex relationships.
- **Gradient Boosting:** Outperformed other models with an R^2 score of **0.986702** and a Mean Squared Error (MSE) of **870.292204**, establishing itself as the most effective method for this dataset.

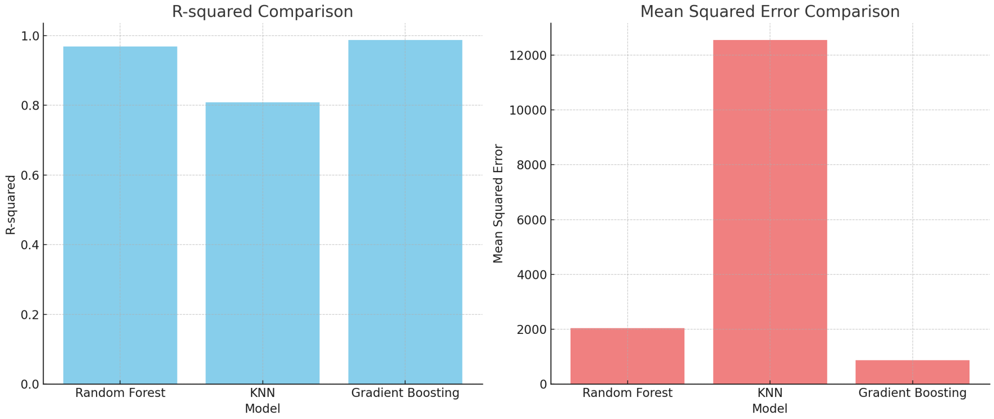

Figure 5: R$\hat{2}$ and mean square error

A comparative analysis of all models was provided, highlighting the most effective model for price predictions according to specifications.

Model	Min	First_Qu	Median	Mean	Third_Qu	Max
Random Forest	2956.594235	3835.569326	6083.172388	35079.738000	11484.204623	151039.149430
KNN	21972.071598	31169.111465	48498.350806	77149.479104	66478.968569	217628.893084
Gradient Boosting	1354.190504	2381.199350	4565.743382	32141.571397	7113.858794	145292.864954

Table 2: Statistical Summary of Cross-Validation Scores for Models

Table 2 provides a statistical summary of the cross-validation scores for three machine learning models: Random Forest, KNN, and Gradient Boosting. The metrics include the minimum (‘Min‘), first quartile (‘First.Qu‘), median, mean, third quartile (‘Third.Qu‘), and maximum (‘Max‘) of the cross-validation scores. These values highlight the performance distribution of each model, with Gradient Boosting showing the lowest minimum and mean error, suggesting its superior performance compared to the other models.

2. Variable Importance:

Feature importance analysis identified the most influential variables for predicting mobile phone prices:

- **processor_speed_ghz** and **ram_gb** were the most significant raw predictors.
- Engineered features like **performance_score** and **price_per_ram** played a critical role in enhancing model interpretability and accuracy.

These insights were visualized through feature importance plots, demonstrating the relative contribution of each variable to the model's predictions.

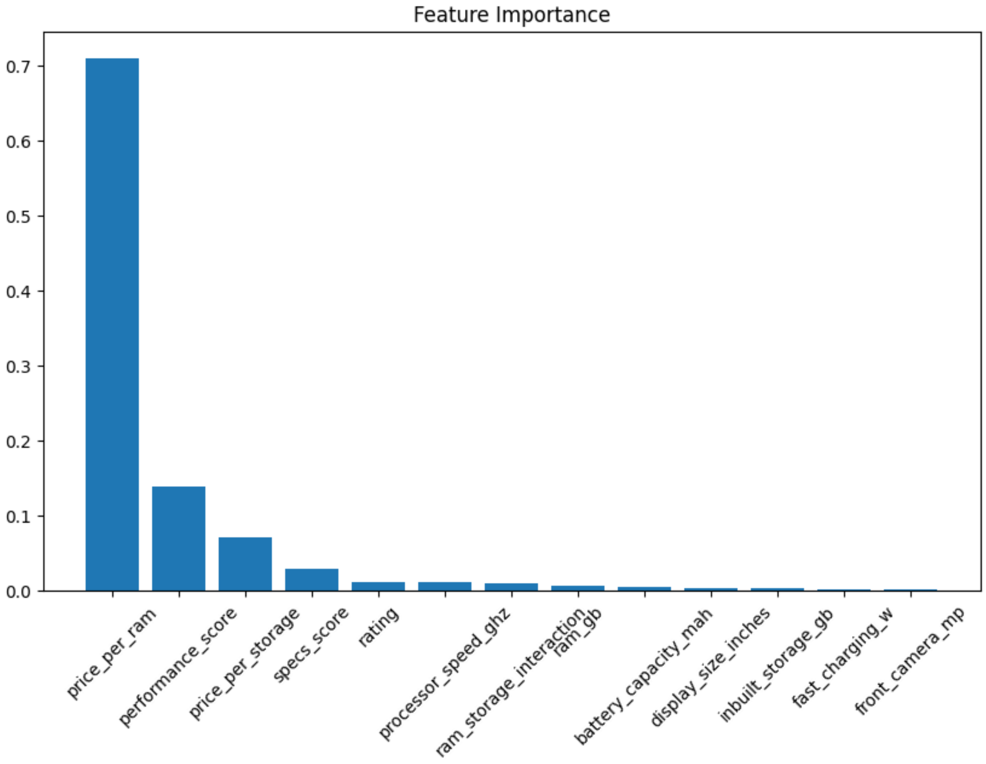

Figure 6: Feature Importance

Figure 5 bar chart ranks the importance of features in predicting the price. price_per_ram is the most significant predictor, followed by performance_score and price_per_storage, while other features have minimal contributions.

3. Prediction Visualization

To further evaluate the predictive power of the Random Forest model, a comparison between actual and predicted prices was visualized. The scatter plot below illustrates the model's performance, with points closely aligned along the diagonal red line, indicating strong predictive accuracy.

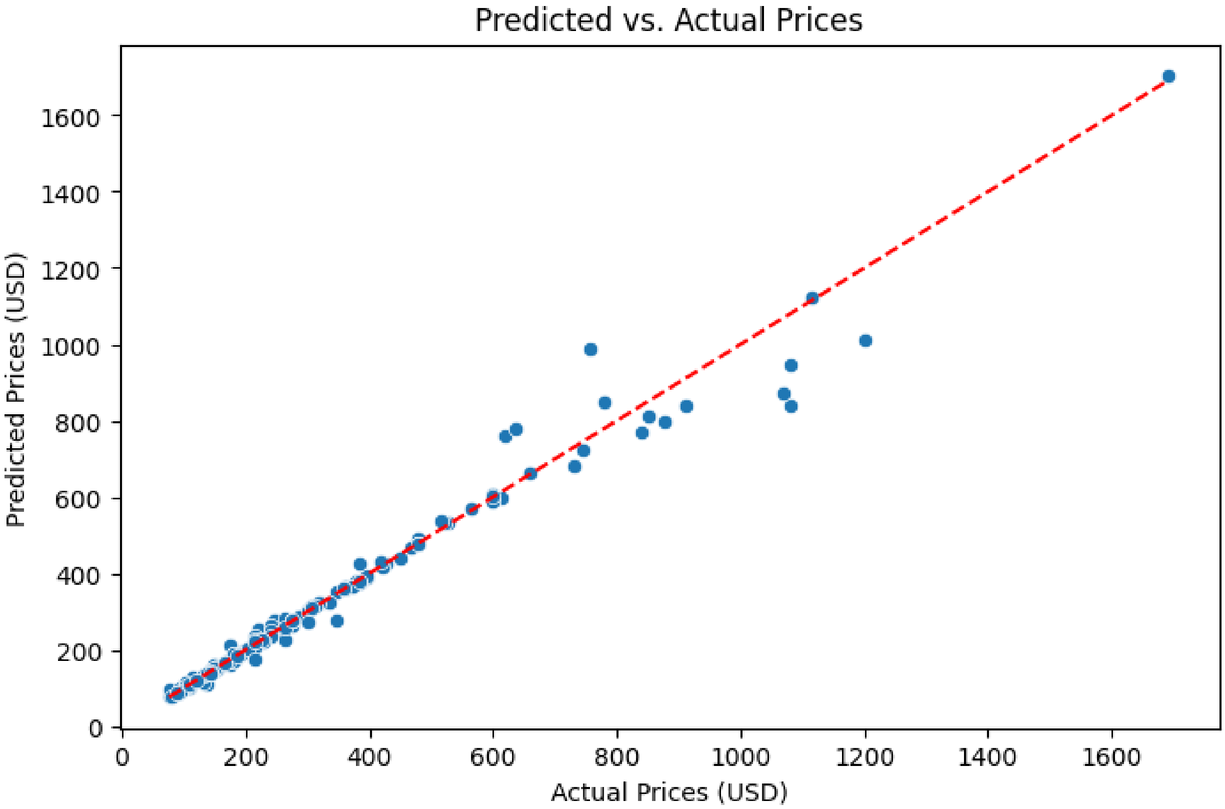

Figure 7: Predicted Prices VS Actual Prices

This visualization underscores the model's ability to predict mobile phone prices effectively, offering confidence in its application to real-world scenarios.

4. Insights from the Model:

The model revealed significant insights into the factors driving mobile phone pricing:

- High-performance metrics such as processor_speed_ghz and performance_score strongly correlate with higher prices.
- Additional features like price_per_ram and battery_capacity_mah offered nuanced understanding of pricing trends.

These findings provide actionable insights into pricing strategies and consumer preferences, contributing to better decision-making in the mobile phone market.

0.7 Conclusion

From the comprehensive analysis using machine learning models, several key conclusions emerge:

- **Predictive Power of Variables:** The study successfully identified crucial variables affecting mobile phone pricing. Features such as processor_speed_ghz, ram_gb, and engineered metrics like performance_score and price_per_ram were found to be the most influential predictors of mobile phone prices.
- **Model Effectiveness:** Among the models tested, the Random Forest Regressor demonstrated the highest predictive accuracy, with an R^2 score of 0.97 and a Mean Absolute Error (MAE) of 16.93. Gradient Boosting also showed strong performance with an R^2 score of 0.98 but slightly higher computational cost.
- **Practical Implications:** The findings provide actionable insights into pricing strategies and consumer preferences in the mobile phone market. By understanding how features interact to influence prices, manufacturers and retailers can make more informed decisions on product development and marketing strategies.

0.8 Limitations

Despite the comprehensive analysis, the following limitations must be acknowledged:

1. **Data Constraints:** The dataset's scope, particularly in terms of diversity and representation of mobile phone models, may limit the generalization of the findings to the broader market. Ensuring a more comprehensive dataset could enhance the applicability of the results.
2. **Model Limitations:** While the Random Forest and Gradient Boosting models demonstrated strong predictive accuracy, their complexity and susceptibility to overfitting remain concerns. The models' reliance on hyperparameter tuning for optimal performance could also limit their scalability in dynamic environments.
3. **Outlier and Missing Value Handling:** While outliers were analyzed and missing values imputed, the choice of median imputation or exclusion may have introduced biases, impacting the accuracy and interpretability of the results. More sophisticated handling approaches might yield better insights.

Given more resources and time, the analysis could be expanded in the following ways:

- **Incorporating Additional Data Sources:** Including a wider variety of mobile phone models across different regions and price ranges could improve the robustness and generalizability of the findings.
- **Consumer Preference Integration:** Integrating data on consumer reviews, preferences, and satisfaction could enrich the analysis, enabling predictions that align more closely with market demands.
- **Deeper Feature Engineering:** Investigating more complex interactions between features, such as polynomial terms or domain-specific metrics, could reveal subtler relationships within the data, further enhancing model accuracy.

0.9 Project Success

The project achieved its primary objective of building an accurate machine learning model to predict mobile phone prices based on key specifications and engineered features. The Random Forest Regressor demonstrated excellent performance, achieving a high R$^{\mathbf{2}}$ score of 0.97 and a Mean Absolute Error (MAE) of 16.93, showcasing its reliability and predictive capability. The analysis successfully identified critical variables, such as ‘processor_speed_ghz‘ and ‘performance_score‘, that influence pricing, offering actionable insights for manufacturers and retailers.

Furthermore, the project effectively tackled challenges like missing data, feature engineering, and model evaluation, highlighting the robustness of the methodology. The insights derived from the study provide a strong foundation for further exploration and practical applications in pricing strategies and market analysis. Overall, the project was a significant success, meeting its goals and delivering valuable outcomes with potential real-world impact.

Bibliography

[1] Waseem Ahmad, Tanvir Ahmed, and B. Ahmad. Pricing of mobile phone attributes at the retail level in a developing country: Hedonic analysis. *Telecommunications Policy*, 2019.

[2] Marcellino Bonamutial and Simeon Yuda Prasetyo. Exploring the impact of feature data normalization and standardization on regression models for smartphone price prediction. In *2023 International Conference on Information Management and Technology (ICIMTech)*, pages 294–298, 2023.

[3] Fanni A. Geyer, Vince A. Szakal, P. A. Kara, and Anikó Simon. Cognitive-bias-induced differences in the perceived video quality of rugged and conventional smartphones. In *2022 16th International Conference on Signal-Image Technology Internet-Based Systems (SITIS)*, pages 592–599, 2022.

[4] Santosh Gupta. Uncleaned mobile dataset, 2024. Accessed: 2024-12-13.

[5] Raju Bhai Manandhar and Jagat Timilsina. Consumer buying decision for smartphones: An analysis of price, brand, and features. *Journal of Nepalese Business Studies*, 2023.