FairPrice: Data-Driven Valuation for Used Smartphones

Group 25

Phyo Zay Thit Andrew Chen 14421871
Ryan Siva 26004313
Darsh Malik 25403546
Yan Hao 25976440
Ritiesh raja 25212477
Jiayu Hao 25948860
Chongyan Tang 26057101

October 5, 2025

Abstract

Used smartphone prices vary widely and often lack transparency, creating uncertainty for both consumers and sellers and slowing the adoption of trade-in programs. Concurrently, electronic waste (e-waste) continues to grow, reaching 62 million metric tons globally in 2022, with approximately 20% attributed to smartphones [1].

This project addresses these challenges by developing a predictive model to estimate the resale value of smartphones using device features, brand, and usage history. A publicly available dataset with specifications and prices for over 3,000 devices was utilized. We employed multiple linear regression with stepwise feature selection, training on log-transformed used prices and inverse-transforming predictions to report performance in the actual price scale.

The model explained 87% of the variance ($R^2 = 0.8689$) and achieved a mean absolute percentage error (MAPE) of 19% with an RMSE of 24.77 on the price scale. Errors were lowest for mainstream brands and medium-RAM devices, while rare brands and niche operating systems exhibited higher error rates due to small sample sizes.

Key findings include:

- A regression model explained 87% of variance and achieved a MAPE of 19% on test data.
- Predictions were most accurate for mainstream Android devices and mid-RAM phones but less reliable for rare brands and smaller OS categories.

The key message is that a simple, explainable regression model can predict used phone prices with strong accuracy. This supports fairer pricing, improves consumer confidence, and contributes to sustainability by encouraging longer device use and reducing e-waste.

Contents

1 Introduction

1.1 Problem statement

Buying a used smartphone can often feel like guesswork — prices vary widely, and many buyers and sellers are unsure if the deal is fair. Consumer behavior research indicates that price uncertainty leads to consumers unwilling to pay, regardless of whether the price favors them [9]. This lack of transparency lowers trust and slows trade-ins. At the same time, smartphones contribute significantly to global e-waste, with an estimated 62 million tons of e-waste generated worldwide in 2022, of which nearly 20% came from smartphones [1]. This project addresses this issue by developing a data-driven model that predicts the fair resale value of used phones, making the process more transparent and sustainable.

1.2 Rationale

Electronic waste is one of the fastest-growing waste streams, driven by rapid technology turnover and consumer demand for new devices. Smartphones are a critical contributor: they are resource-intensive to produce, contain rare earth elements, and depreciate quickly in resale markets [10]. When undervalued, many phones are discarded prematurely even though they remain functional, adding to landfill pressure and wasting valuable materials.

Better valuation can address multiple issues:

1. Consumer trust. Buyers and sellers need consistent benchmarks to negotiate fair deals. According to [3], higher confidence in the valuation leads to the customer's willingness to pay.
2. Business efficiency. Trade-in programs and e-commerce platforms benefit from automated fair pricing.
3. Sustainability. Longer use extends device lifecycles, reducing e-waste and resource waste.

Previous research highlights three key factors in smartphone resale value. First, time since release is the strongest driver of depreciation [10]. Second, brand plays an important role, with premium brands such as Apple and Samsung showing slower depreciation [5]. Third, technical specifications such as RAM, storage, and battery capacity can moderate depreciation [2]. Understanding these literature reviews provides confidence that there is signal to be found in these features, which may result in an accurate model.

Recent studies also compare modeling approaches. For instance, Random Forests outperform linear models in some price prediction tasks [7], while others stress the need to provide price intervals instead of point estimates to better manage market risks [4]. These insights guided our research design: build a transparent regression model as a baseline, evaluate predictive performance, and discuss paths toward more advanced approaches.

1.3 Project aims and objectives

The aim of this research is to develop a predictive model that accurately estimates the resale price of used smartphones based on their features, usage history, and brand, thereby supporting fairer pricing decisions and identifies the main drivers of depreciation.

To support our project aim, we asked the following questions:

1. Does usage time (days used) significantly reduce resale price?
2. Do technical specifications such as RAM and battery predict resale price?
3. Does brand and OS explain differences in depreciation beyond technical specifications?
4. Can a stepwise multiple linear regression predict resale price with acceptable error on a held-out test set?

To answer these questions, we had the following objectives:

1. Collect and explore a real-world dataset of smartphones with 13 features and resale prices.
2. Preprocess and clean the data to handle missing values and outliers.
3. Conduct exploratory data analysis to identify relevant predictors and check for collinearity.
4. Train and evaluate regression models with stepwise feature selection.
5. Inverse-transform predictions and report business metrics such as MAPE.
6. Evaluate subgroup accuracy by brand, OS, RAM, and usage.
7. Interpret results in the context of market fairness and sustainability.

In short, we show that age, usage duration, and device specifications are the strongest predictors of second-hand value, offering important insights for buyers, sellers, and businesses in the smartphone ecosystem.

2 Methodology

2.1 Methods Overview

The project followed a structured process, progressing from data understanding to model development. A real-world smartphone dataset was collected, containing device specifications (RAM, battery, release year, brand) and usage time. The data were prepared and cleaned by addressing missing values and unrealistic entries.

Exploratory analysis was conducted to identify patterns (e.g., whether newer models or larger batteries retain value better). Finally, predictive models were developed using regression methods, with evaluations focusing on both statistical accuracy and practical relevance for resale markets.

2.2 Methods Details

2.2.1 Data Acquisition

A publicly available dataset ($n \approx 3{,}454$) was used, containing resale prices and features of smartphones. The dataset included 13 features per device:

- Brand and operating system (OS)
- Screen size, RAM, and storage capacity
- Battery capacity
- Front and back camera specifications
- 4G/5G support
- Days used and release year
- Normalized (log-transformed) resale price (target variable)
- Normalized (log-transformed) retail price

The target variable (normalized resale price) enabled fair comparison across different brands and models.

2.2.2 Exploratory Data Analysis (EDA)

EDA was conducted in Python using `pandas` and `matplotlib/seaborn`. The following visualizations were generated:

- Scatter plots to examine relationships between resale price and usage time
- Heatmaps to explore correlations among technical specifications
- Boxplots to visualize price differences across brands

These visualizations refined the selection of candidate predictors for the regression model. Detailed patterns and numerical relationships are presented in the Results section. EDA also confirmed significant correlations between the target variable and features, indicating sufficient signal for model development.

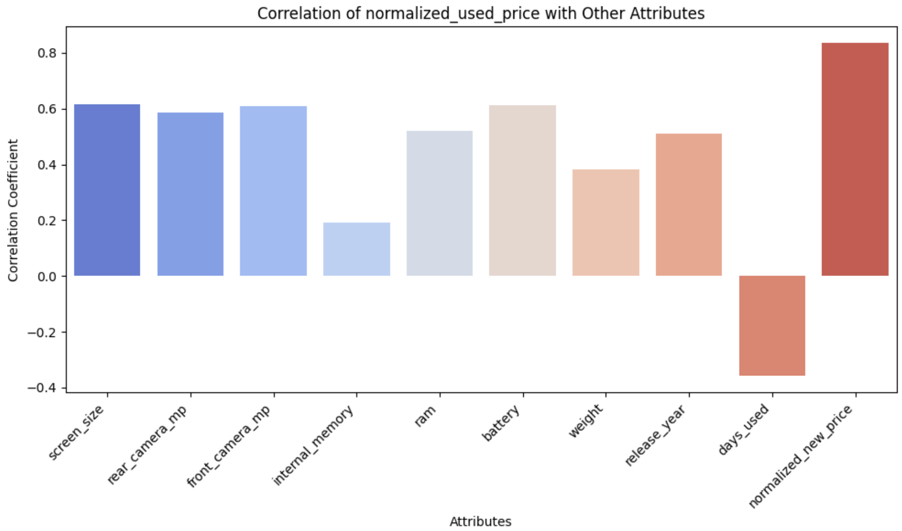

Figure 1: Correlation of normalized used price with Other Attributes

2.2.3 Data Preparation and Cleaning

Data preparation involved three key steps:

1. **Missing Values**: Records with incomplete feature information (e.g., missing battery capacity) were either imputed or excluded, depending on the severity of missingness.
2. **Outliers**: Implausible values (e.g., devices listed as used for over 4,000 days or camera megapixels outside realistic ranges) were treated to avoid distorting results.
3. **Distribution/Skewed Data**: Prices labeled as "normalized" were verified to be natural log-transformed values. After cleaning, 3,253 records remained for analysis.

2.2.4 Modeling Approach

An explainable linear model with stepwise feature selection was selected to balance accuracy, interpretability, and operational use. Linear models provide transparent coefficients for audit and pricing policy—critical for consumer-facing markets. Stepwise selection reduces redundant predictors and mitigates collinearity identified during EDA. While non-linear models (e.g., Random Forest or gradient boosting) may improve accuracy, linear models are easier to govern and deploy with clear price justifications [5, 10].

2.2.5 Assumptions and Diagnostics

Standard linear regression assumptions were validated through the following checks:
1. **Normality**: Residual plots and Q-Q plots were examined. Although features were not strictly normal, they exhibited approximate normality. The target variable (normalized resale price) was confirmed to follow a normal distribution.

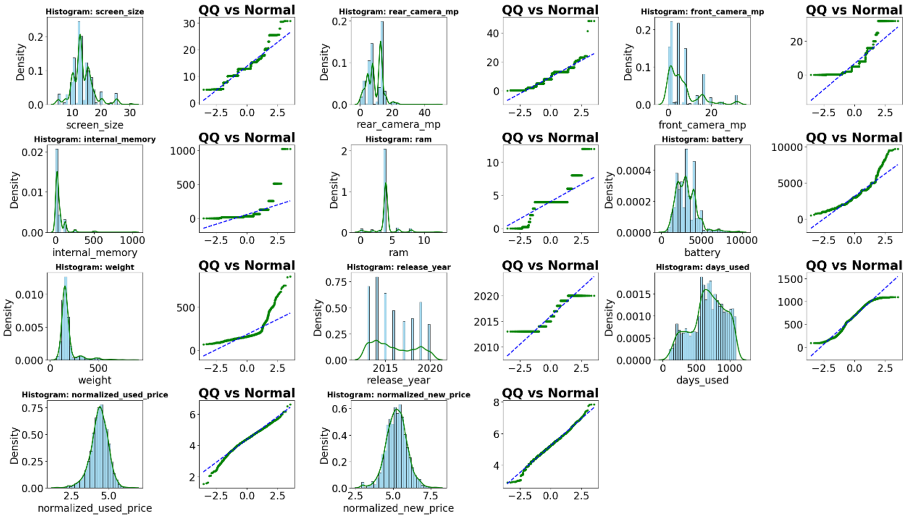

Figure 2

2. **Linearity**: Scatter plots verified that most features maintained linear relationships with the target variable.

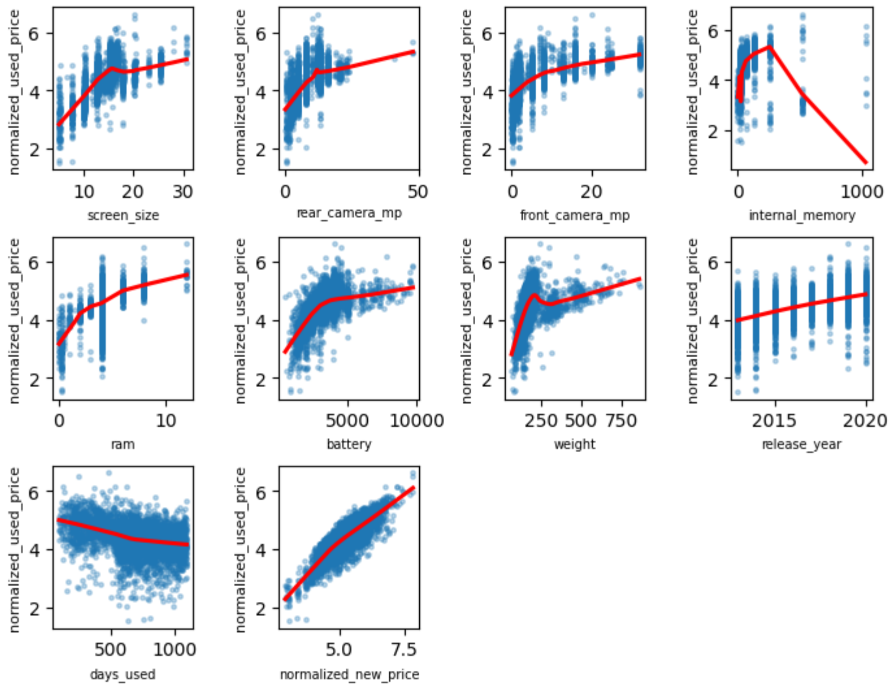

Figure 3

3. **Correlation**: A correlation matrix revealed significant multicollinearity between features (e.g., screen size, battery capacity, and weight). This was mitigated through stepwise variable selection.

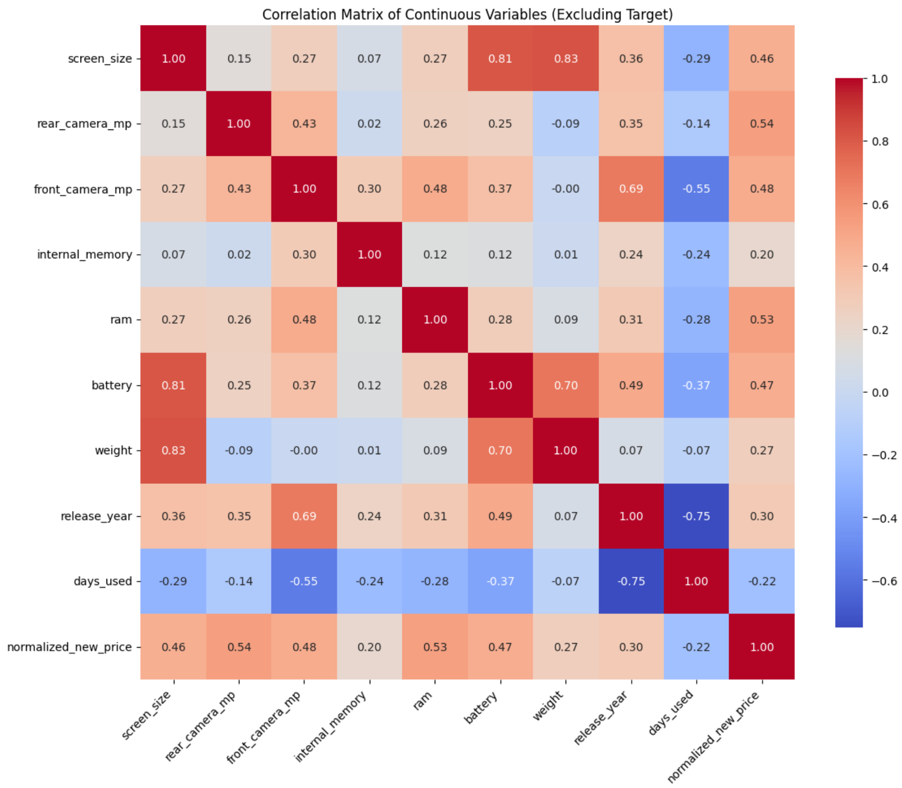

Figure 4

4. **Heteroskedasticity**: Robust standard errors were applied to address potential heteroskedasticity. Residual vs. fitted values plots (post-prediction) were used to confirm this assumption.

2.2.6 Model Evaluation

The model was evaluated on the held-out test set (80/20 split). Performance metrics included:

- R^2 and RMSE in the log-transformed space
- RMSE and MAPE after converting predictions back to the original price scale
- Subgroup error rates (by brand, OS, RAM category, and usage duration) to assess accuracy across dataset segments

2.2.7 Limitations in Methodology

The dataset had several limitations:

- Absence of software support information or repair history (known to affect real-world resale pricing)
- Brand imbalance (overrepresentation of Apple devices)
- Lack of condition-related features (e.g., scratches, battery health), reducing generalizability

The model was evaluated on a single public dataset and time window. For deployment across markets or future periods, periodic temporal validation and re-calibration using recent sales data are recommended to maintain accuracy.

2.2.8 Ethical Considerations

When developing and deploying the used smartphone valuation model, several ethical considerations must be addressed to ensure fairness, privacy, and reliability for all stakeholders. These considerations—along with their potential impacts and mitigation intentions—are detailed below:

1. **PII and Data Privacy** This is a critical ethical priority, as mishandling personal data poses substantial risks to user privacy. For the current study, we have thoroughly validated the dataset and confirmed that none of the included features contain personally identifiable information (PII) related to buyers or sellers (e.g., names, contact details, device serial numbers linked to individuals). This validation ensures compliance with data protection principles and mitigates the risk of unintended privacy breaches.

2. **Brand Bias** The dataset exhibits a significant imbalance in brand representation: mainstream brands (e.g., Apple, Samsung) are overrepresented, while niche or regional brands (e.g., Google Pixel, Infinix) are underrepresented. This imbalance may lead to unfair and inaccurate price predictions for underrepresented brands, as the model lacks sufficient data to learn their unique pricing patterns. We have explicitly noted this issue, and prior to any production deployment, we will require supplementary data sampling (e.g., stratified sampling of underrepresented brands) to balance the dataset and reduce prediction bias.

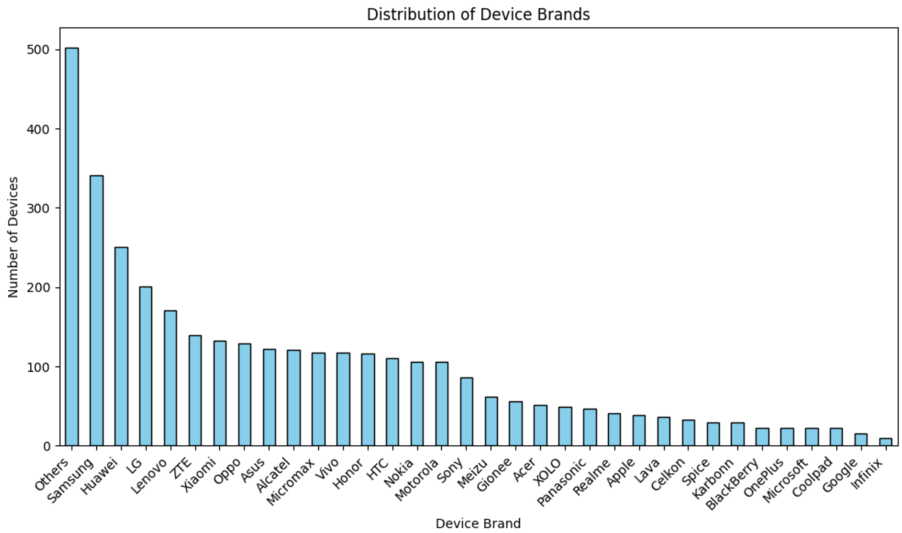

Figure 5: Distribution of Device Brands

3. **Marketing Segment Bias** The origin of the dataset (e.g., specific online platforms, regional markets, or demographic groups) is not clearly documented. This ambiguity introduces the risk of marketing segment bias: the data may reflect pricing behaviors

specific to a narrow subset of the market (e.g., users of a single resale platform) rather than the broader used smartphone market. For example, prices in the dataset might include platform-specific fees (e.g., commission charges) that do not apply to offline transactions or other platforms, limiting the model's generalizability across different market segments.

4. **Economic Environment Sensitivity** Over time, the model may fail to generalize effectively because it does not account for macroeconomic factors that significantly impact consumer willingness to pay (e.g., inflation, currency fluctuations, economic recessions). These factors are not included in the current feature set, so as economic conditions change, the model's predictions may become less accurate (a form of "model depreciation"). To address this, we will monitor the model's performance over time; if we identify declining accuracy due to economic shifts, we will either retrain the model with updated data or introduce new features to capture relevant economic environment factors.

5. **Transparency and Explainability** A key ethical principle for consumer-facing pricing tools is transparency. The current model uses a clear, interpretable set of features (e.g., RAM, days used, brand, battery capacity), which allows for straightforward explanation of price predictions. This means customers can easily understand the factors driving their device's estimated value—for instance, why a phone with more RAM or fewer days of use is valued higher. This level of explainability aligns with ethical AI guidelines that prioritize user trust through clear communication of model logic [6].

6. **Historical Bias** The model relies on historical pricing data to learn depreciation patterns, which introduces the risk of historical bias. As market preferences shift (e.g., consumers begin favoring new brands, or a brand's product quality improves over time), the model may not capture these changes. Instead, it may continue to underpredict the value of emerging or improved brands based on their past performance, leading to unfair valuation for sellers of these devices. This bias highlights the need for regular updates to the dataset to reflect current market dynamics, ensuring the model's predictions remain aligned with real-time consumer preferences.

3 Results

3.1 Key Findings

The stepwise regression model predicted used smartphone prices with strong accuracy. On the test set, the model explained 87% of variance ($R^2 = 0.87$) and achieved a MAPE of 19% on the actual price scale, with a corresponding RMSE of 24.77.

Errors were lowest for mainstream brands, Android devices, and medium-RAM phones, while higher errors occurred for rare brands, high-RAM devices, and those with niche operating systems.

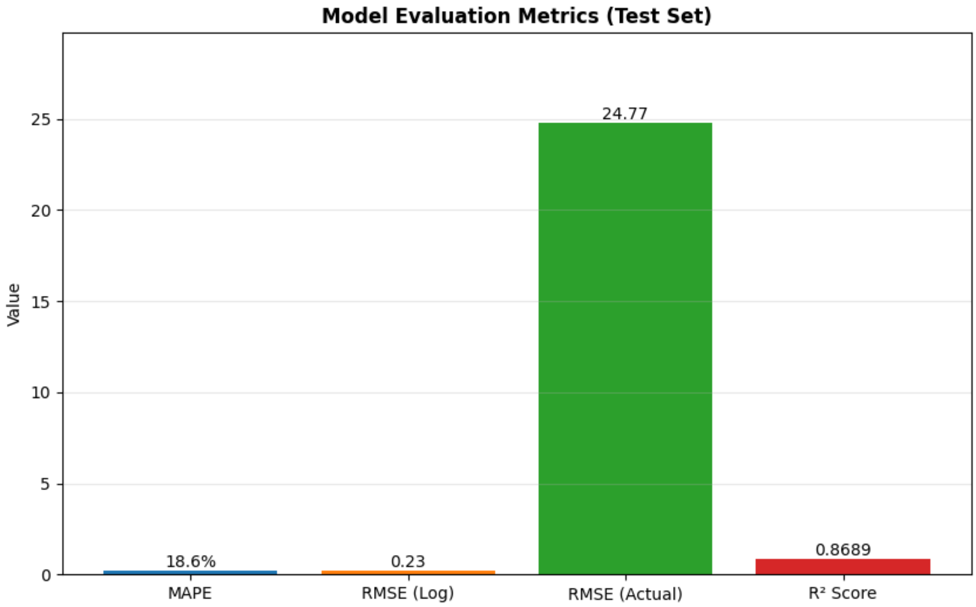

Figure 6: Overall model accuracy metrics on the test set

3.2 In-depth Results

3.2.1 Does usage time (days used) significantly reduce resale price?

Resale price declined significantly with increased usage time. A correlation heatmap indicated a negative association ($r = -0.36$), confirming depreciation over time. Scatter plots showed that the steepest price decline occurred in the first 200–300 days of use.

Model subgroup analysis showed consistent error levels: lightly used phones ($\leq$1 year) had MAPE = 0.18, while moderately used phones (1–3 years) had MAPE = 0.19. This indicates that the model captured depreciation patterns effectively across different usage durations.

These findings are consistent with prior research showing that age is the dominant factor in smartphone value loss [2, 10]. Similar evidence from hedonic pricing models confirms that consumer willingness-to-pay drops quickly with device age [8].

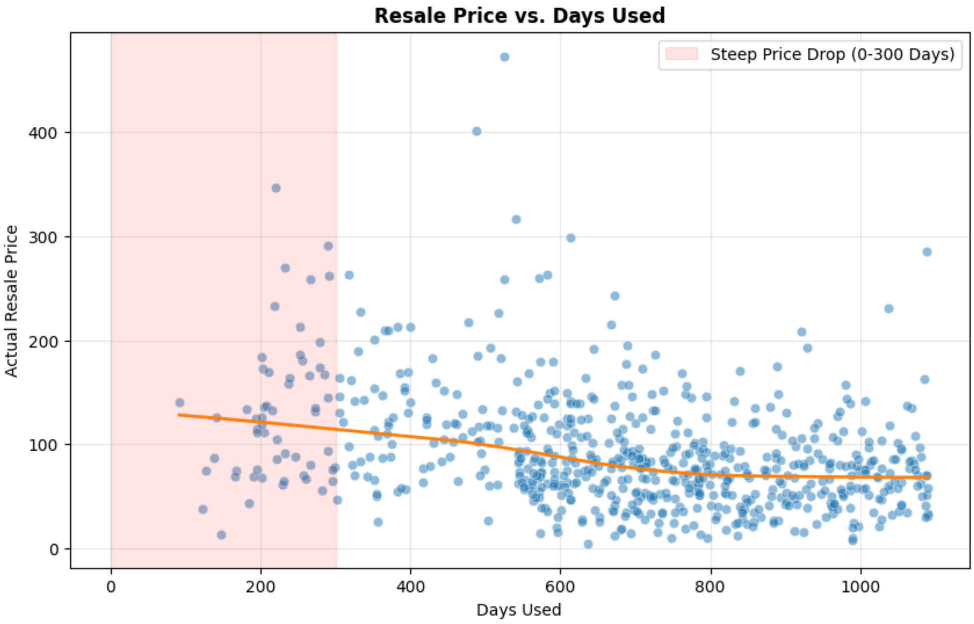

Figure 7: Scatter plot of resale price vs. days used

3.2.2 Do technical specifications such as RAM and battery predict resale price?

Technical specifications were moderately associated with resale value. RAM showed a positive correlation ($r = 0.52$). Medium RAM devices (3–4GB) had the lowest error (MAPE = 0.18), while low RAM devices ($\leq$3GB) performed worse (MAPE = 0.21). High RAM devices (>4GB) showed the largest error (MAPE = 0.24), likely due to smaller sample sizes.

Battery capacity also correlated positively with resale price ($r = 0.61$). Devices with larger batteries tended to retain value better, aligning with buyer preferences for longer battery life.

These results are consistent with prior findings that technical specifications, particularly memory and battery capacity, are significant but secondary drivers of resale value compared to age and brand [2, 10].

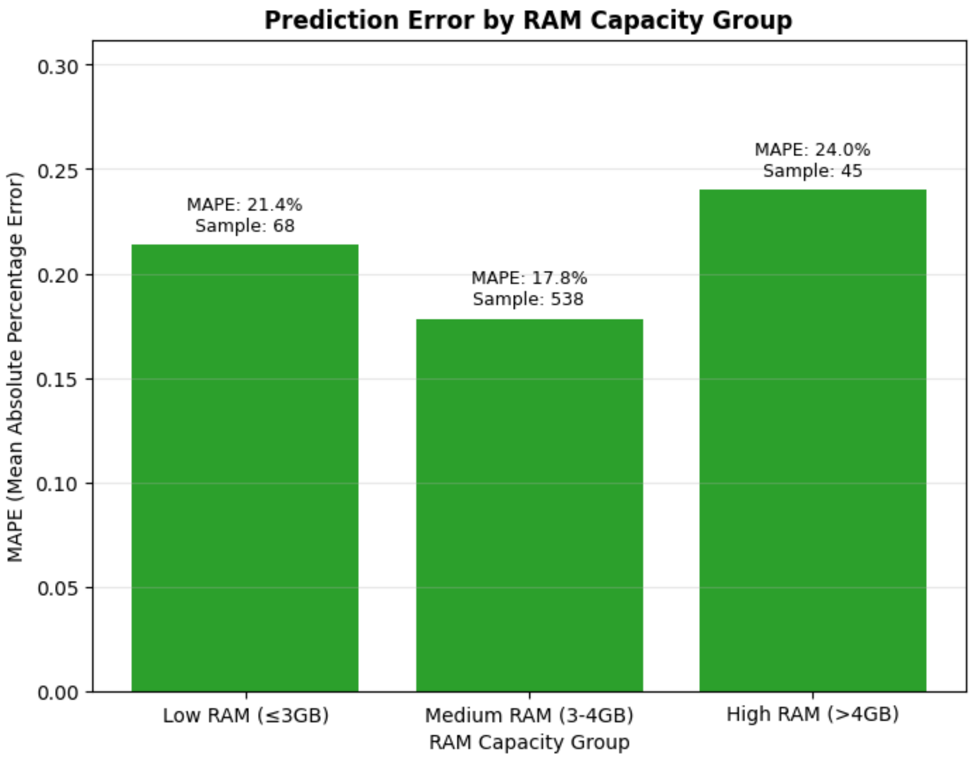

Figure 8: MAPE error rates by RAM capacity groups

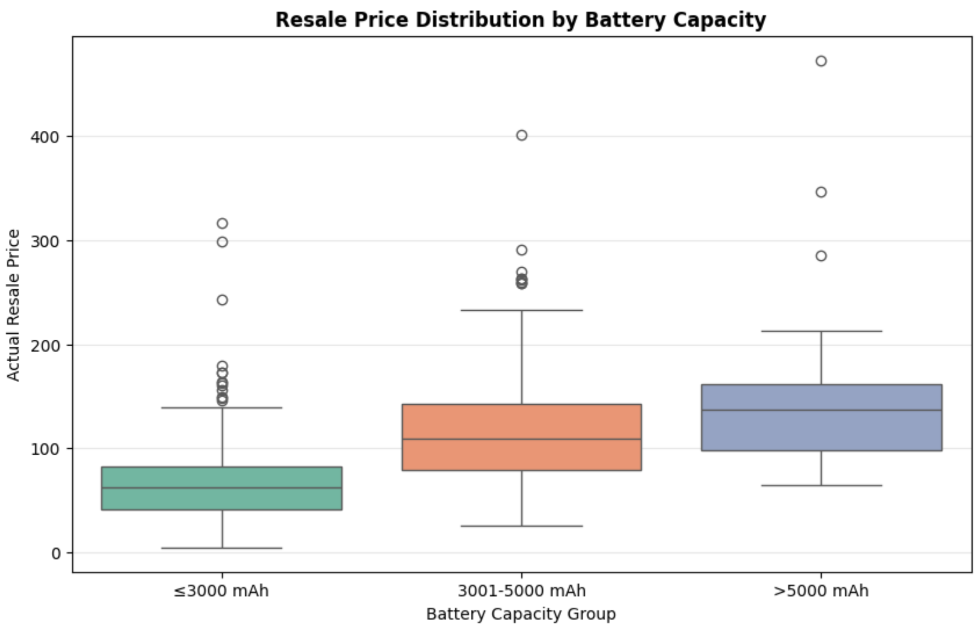

Figure 9: Distribution of resale price by battery capacity

3.2.3 Does brand and OS explain differences in depreciation beyond technical specifications?

Brand and operating system both had clear effects on resale value beyond technical specifications.

Brand. Major brand groups (Samsung, Huawei, Lenovo) showed stable performance (MAPE $\approx$ 0.18–0.19). Premium brands such as Apple and Samsung retained value longer, consistent with evidence from second-hand marketplaces [5]. In contrast, rare brands (Group 0, $n = 8$) had much higher error (MAPE = 0.40), reflecting limited representation in the dataset.

Operating system. Depreciation also varied by OS:

- Android devices ($n = 596$): MAPE = 0.18
- Windows devices ($n = 13$): MAPE = 0.16
- iOS devices ($n = 9$): MAPE = 0.20
- Other OS devices ($n = 33$): MAPE = 0.32

Boxplots highlighted that Android and iOS devices retained value more consistently than smaller OS groups. These findings align with prior studies suggesting that longer software update support and ecosystem stability reduce depreciation [8, 10].

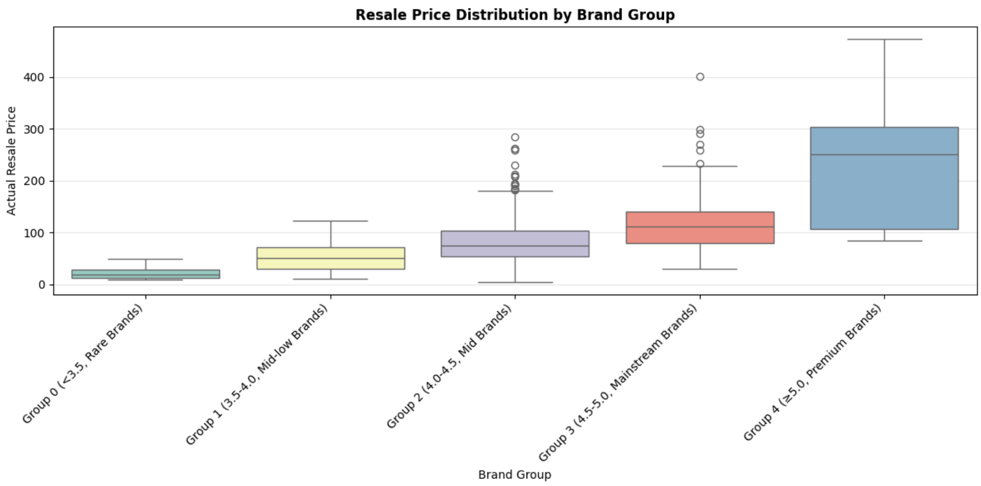

Figure 10: Boxplot of resale value by brand group

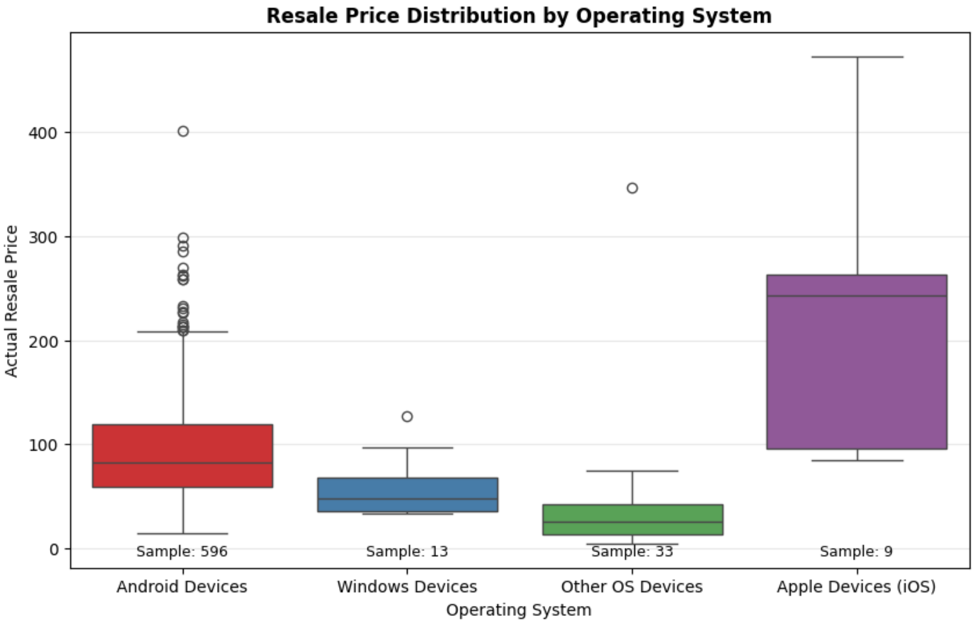

Figure 11: Boxplot of resale value by operating system

3.2.4 Can a stepwise multiple linear regression predict resale price with acceptable error on a held-out test set?

The stepwise regression model performed strongly in predicting log-transformed resale prices and in translating results back to the price scale. On the test set:

- $R^2 = 0.8689$ (log-transformed scale)
- RMSE = 0.23 (log scale)
- RMSE = 24.77 (price scale)
- MAPE = 19% (price scale)

These metrics indicate that the model has practical accuracy for real-world trade-in and resale markets, where a 20% margin of error is generally acceptable for pricing tools [4].

Linear regression assumptions were checked: residual plots and Q-Q plots showed no major violations, and robust standard errors were applied to address minor heteroskedasticity.

The findings are consistent with previous research showing that regression-based models can explain the majority of variance in smartphone resale prices, though non-linear models such as Random Forests may further improve accuracy [7].

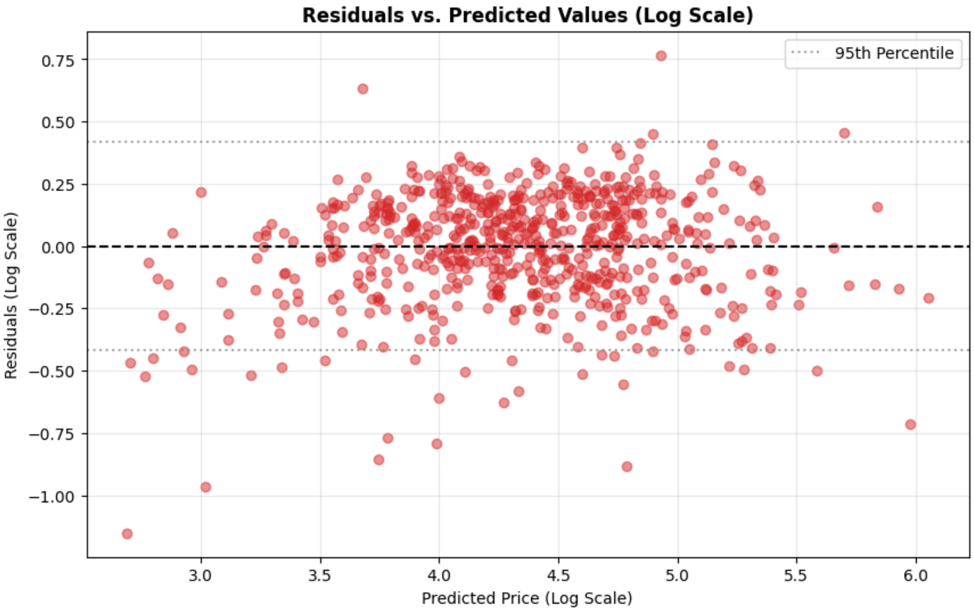

Figure 12: Residual plot of the stepwise regression model

3.2.5 Outliers and Reliability

Prediction reliability decreased for categories with sparse data, such as rare brands and "Other OS" devices, where MAPE exceeded 30%. These results highlight the importance of sufficient data coverage for accurate predictions.

3.2.6 Summary

Our results show that usage time and brand/OS are the strongest drivers of resale price, with RAM and battery capacity providing moderate contributions. The regression model achieved $R^2 = 0.87$ and MAPE = 19%, supporting its use as a baseline predictive tool for resale markets. These results align with earlier evidence on smartphone depreciation and extend the case for predictive analytics in second-hand device pricing [2, 5, 10].

4 Conclusion

4.1 Take home message

The core takeaway from this research is that a multiple linear regression model, constructed through stepwise feature selection, can accurately predict the resale price of used mobile phones. It explains 87% of price variation ($R^2 = 0.8689$) and achieves a mean absolute percentage error (MAPE) of 19% on the actual price scale. Key influencing factors encompass normalized new price, release year, screen size, camera specifications (front/rear camera megapixels), RAM, weight, 4G support, and days used.

This finding directly tackles the ambiguity in second-hand phone pricing, a long-standing issue that has impeded consumer trust and the adoption of trade-in programs. By offering transparent, data-driven price estimates, the model facilitates fairer transactions and aligns with broader sustainability objectives—promoting longer device lifecycles to reduce electronic waste.

Future work could expand on this by integrating real-world features (such as device wear, repair history, and software support) and exploring advanced models (like gradient boosting or neural networks) to minimize prediction errors for rare brands, thus enhancing the model's applicability across the entire used phone market spectrum.

4.2 Discussion

The primary goal of this project was to build a statistical model for predicting used mobile phone resale prices and identifying the key factors driving price fluctuations—and this objective was fully accomplished. The model's performance attests to this: an R^2 value of 0.8689 shows that nearly 87% of the variation in used phone prices can be explained by the selected features, while a 19% MAPE on the actual price scale confirms the model's predictions are sufficiently accurate for practical use. The operational significance of MAPE lies in its ability to minimize over- or under-pricing and curb consumer dissatisfaction, which is vital for fostering trust in pricing mechanisms within the second-hand mobile phone market [3].

When contextualized within the wider landscape of second-hand electronics markets, this result is consistent with prior research, which established that technical specifications and usage duration are primary drivers of device depreciation. Our results further reinforce three key patterns well-documented in the literature:

1. Depreciation is dominated by time since release or use [10];
2. Brand effects are persistent and substantial in secondary markets [5];
3. Technical specifications like memory and battery add value but have smaller average effects compared to time and brand [2].

Operationally, prediction tools gain from using price bands instead of point estimates to reduce negotiation friction and mispricing risk [4]. For scenarios requiring higher accuracy, non-linear models can supplement the linear baseline [7]. In the context of the circular economy, transparent pricing is associated with increased participation in reuse and take-back programs [1, 8].

A notable finding in our research was the significant difference in prediction accuracy across brand groups: premium and mainstream brands had the lowest MAPE (0.17, $n = 192$), whereas rare brands had a much higher error (0.40, $n = 8$). This disparity is likely due to the extremely small sample size of rare brands, which restricts the model's capacity to learn their unique pricing patterns.

Beyond used phones, these findings have wider implications for the second-hand electronics market: the framework of "core technical specs + usage time + brand value" could act as a generalizable model for predicting resale prices of other consumer electronics, aiding in the standardization of pricing across fragmented second-hand markets.

Collinearity was detected in Exploratory Data Analysis (EDA) and mitigated through stepwise selection, which helped alleviate issues related to feature intercorrelations and improved the model's interpretability.

4.3 Project limitations and caveats

This research has several limitations that should be considered when interpreting its results:

1. **Small sample sizes for niche brands**: Some brand groups (e.g., Brand Group 0) had only 8 samples, leading to disproportionately high prediction errors (MAPE = 0.40). This implies the model cannot reliably represent the pricing behavior of rare or regional brands, limiting its applicability for users looking to value less common devices.

2. **Missing critical real-world features**: The dataset lacked variables that strongly influence used phone prices in practice, such as device wear (e.g., scratches, battery health), repair history (e.g., whether the screen was replaced), or software support (e.g., remaining years of OS updates). Omitting these features likely underestimated their impact on resale value and might have introduced bias into the model.

3. **Multicollinearity issues**: Collinearity was identified in EDA and reduced through stepwise selection. However, some intercorrelations between features may still exist, which could have distorted the coefficients of correlated features to some degree, reducing the reliability of individual feature impact assessments.

4. **Standardized price limitations**: The dataset used "normalized" prices rather than raw currency values, necessitating conversion for real-world application. This conversion process could introduce minor errors, and the unclear normalization method (e.g., whether it accounts for regional price differences) limits the model's usability in global markets.

5. **Ethical and IGA considerations**: We utilized public, de-identified data and followed data-minimization principles. If future versions include user-level or location attributes, we will obtain informed consent and respect data sovereignty and custodianship. To reduce fairness risks, we will monitor subgroup error (by brand, OS, price tiers) and direct low-confidence predictions to human review. These measures support culturally safe practice and responsible AI use in market pricing.

4.4 Stakeholder analysis and project outcomes

4.4.1 Key Project Outcomes

This research yields two actionable outcomes:

1. A ready-to-deploy multiple linear regression model capable of generating accurate resale price estimates for used phones using easily accessible features (e.g., specifications, usage time, brand).

2. A clear ranking of factors influencing resale value—with normalized new price and release year as the strongest drivers, followed by hardware specs (RAM, camera) and usage duration—providing evidence on how to maximize or assess a used phone's value.

4.4.2 Relevant Stakeholders and Benefits

1. **Second-Hand E-Commerce Platforms**: Can use the predictive model to automate "fair price" recommendations for sellers, reducing manual effort and standardizing pricing across listings.

2. **Used Phone Sellers**: Can leverage the pricing knowledge to prioritize value-preserving actions (such as choosing brands like Apple or Samsung, minimizing usage time, or highlighting high RAM and camera specs) to maximize resale revenue.

3. **Consumers**: Can use the model to verify if a listed price is reasonable, avoiding overpayment for devices with poor value retention (e.g., niche brands or heavily used phones).

4. **Mobile Phone Manufacturers**: Can apply insights into "value-preserving features" (e.g., high-quality cameras, sufficient RAM) to guide product design, helping improve the long-term resale value of their devices and enhance brand loyalty.

References

[1] V. Forti, C. P. Baldé, R. Kuehr, R. J. Cox, J. Huisman, F. Wang, D. Tetteh, and B. Saygın. The global e-waste monitor 2024: Tackling the e-waste challenge, 2024.

[2] JST. Used smartphone: An analysis of factors affecting its price. *Journal of the Japanese Society for Management Information*, 2018s:108–113, 2018.

[3] J. Kas, R. Corten, and A. van de Rijt. Trust, reputation, and the value of promises in online auctions of used goods. *Rationality and Society*, 2023.

[4] R. Kumar, V. Singh, and A. Sharma. Estimating smartphone price ranges using machine learning models. In *Advances in Intelligent Systems and Computing*, pages 945–954. Springer, 2023.

[5] T. Makov, T. Fishman, M. R. Chertow, and V. Blass. What affects the secondhand value of smartphones: Evidence from ebay. *Journal of Industrial Ecology*, 23(3):549–559, 2019.

[6] S. Pandey. Privacy and ethical considerations in ai. *LinkedIn*, 2024. Published online 5 March.

[7] V. Sarayu and S. Vidyavathi. Predicting smartphone prices using machine learning algorithms. *Advances in Computational Research*, 13(2):57–64, 2021.

[8] B. Shen, Z. Yu, and L. Zhou. The service strategy and influencing factors of online recycling of used mobile phones. *Journal of Cleaner Production*, 281:125327, 2021.

[9] C. Tucker. Pricing under consumer uncertainty. `https://ocw.mit.edu/courses/15-818-pricing-spring-2010/resources/mit15_818s10_lec03/`, 2010. Lecture notes.

[10] L. Zhou and S. M. Gupta. Value depreciation factors for new and remanufactured high-technology products: A case study on iphones and ipads. *International Journal of Production Research*, 58(23):7218–7249, 2020.