**USED SMART PHONE PRICE ANALYTICS**

**Phanidhar Venkata Naga Kasuba**

**CSDA 6010 Z1 FA – Analytics Practicum**

**Webster University**

**Professor: Dr. Willie Rivers**

**Executive Summary**

The main issue that this project has solved is the issue of correctly estimating the resale value of used smartphones in a fast-changing resale market. Unlike new gadgets with set prices by the manufacturers, the prices of used smartphones are determined by various interdependent factors which include brand reputation, operating system, hardware specifications, age, and network compatibility. In the absence of evidence-based estimates, sellers may be at risk of undervaluing devices and missing potential revenues, and buyers may not be sure whether they are paying fair prices. Such a lack of transparency lowers profitability, affects competitiveness, and erodes consumer trust, which is why analytical models that can be used to predict and classify resale prices are necessary.

So far, much progress has been achieved by understanding and exploratory data analysis (EDA). The data set with 3,454 devices and 15 attributes have been analyzed in terms of missing values, zeros, distributions and patterns. Categorical insights demonstrate that Android is the leader in the market, 4G is a common technology, and 5G is not very common, and the distribution of the brands is not balanced, with the highest share belonging to the category Others. The numerical analysis shows that there is a tendency to cluster around average specifications like 4 GB RAM, 16-64 GB internal storage and 2000-4000 mAh battery, with outliers observed in screen size, memory and battery. Both standardized old and new prices are bell-shaped, which indicates that there are consistent ranges and depreciation patterns. Collectively, these results form a solid basis of the further steps of predictor analysis, data transformation and model development to meet the business and analytical objectives of the project.

**Introduction**

The secondary handset market has become an important part of the overall mobile business, as the demand for consumers to buy low-cost devices is rising, technology is evolving at a fast rate, and the release of new models is frequent. With more consumers turning to pre-owned devices to save money, be more environmentally friendly, or have access to higher-end models at a lower cost, the pricing dynamics of such devices have become of greater importance. In contrast to new phones, where price is set by the manufacturer and retailer, used smartphone prices are affected by many different variables including age, features, condition, and overall market demand.

The current project explores the connections between smartphone characteristics and the used phone resale value. By comparing the features like screen size, RAM, internal storage, battery life, camera specs, and compatibility with technologies like 4G and 5G, this project will reveal which features are the most influential in deciding the value of used phones. Exploratory data analysis (EDA) will be employed in order to analyze the dataset and determine the patterns, correlations, and potential outliers. This step will also show any multicollinearity between variables that can affect the predictive ability of the machine learning models.

The analytical model of this project combines regression and classification. Regression will be useful in predicting continuous variables, like the normalized resale price, whereas classification models will enable the division of devices into categories like high price and low price. This twofold strategy will mean that the project not only forecasts precise numbers but also comes up with general categories that will be helpful to both sellers and buyers in coming up with informed decisions. In so doing, the study turns the dataset into clear, practical guidance for sellers and buyers, so that market participants can better comprehend the trade-offs between features and resale value of smartphones.

Lastly, the project shows the value of data-driven decision-making in a competitive and dynamic industry. The project can be used to advance the academic knowledge of predictors of resale value and create robust machine learning models and test their efficiency. The results will not only point out the fundamental factors that influence the prices of used smartphones but will also form a foundation on which businesses, resellers, and consumers can make more informed decisions in the market.

**Business Goals and Problems**

**Business Goals**

The prices of used phones are commonly determined by guesswork despite value being based on brand, operating system, specifications (RAM, storage, cameras, battery, screen size), age, and 4G/5G-this causes profit loss on high-end phones, sluggish inventory turnover and old stock, inconsistent trade-in offers, and customer mistrust; we must have a transparent, data-driven method of setting fair prices that will sell more inventory without losing profit.

**Business Problems**

The prices of used phones are usually determined through guesswork despite the fact that the value is determined by brand operating system hardware specifications age and network capability. This leads to missed profit on high-end devices sluggish inventory turns lopsided trade ins and lack of customer confidence. The company requires an open data driven approach to establish equitable prices that safeguard margin accelerate sales and maintain listings uniformity across brands and circumstances.

**Analytical Goals and Approach**

**Analytical Goals**

Provide a low error price predictor and a high accuracy tiering classifier with equal sensitivity and specificity. Give clear explanations behind every quote with model agnostic attributions to ensure that sellers and buyers understand the reasoning behind a price proposal. Make sure that probabilities are well calibrated to ensure that predicted levels are observed. Be stable within brands and operating systems and over time and space. Show resistance to outliers and missing fields and moderate data drift. Deliver practical diagnostics that indicate the levers that change price like memory camera battery screen size network capability and device age. Measure uncertainty using prediction intervals to allow teams to establish guardrails on discounts and trade ins. Record data lineage and modeling decisions and publish straightforward visuals that can be easily understood by stakeholders.

**Analytical Approach**

Cleanse the data and correct data quality problems with type corrections and missing value processing and sensitive outlier treatment based on domain rules. Extract interpretable values like age of device based on release year and days used and a value retention indicative based on the relationship between used price and new price and network capability and brand segments flags. Categorical fields should be encoded, numeric fields scaled where necessary and multicollinearity should be checked and redundant predictors merged or dropped where they decrease the stability. Divide the data into training and testing and repeat cross validate on selection and tuning. The linear and stepwise models are then fitted as transparent baselines and ensembles like random forest and gradient boosting are then trained to capture interactions and nonlinear effects. Calibrate classifier results and add per record descriptions with shap or other methods. Conduct a controlled pilot in the current pricing process whilst tracking the conversion time to sell return rate and margin influence. Add fairness checks by brand operating system and network capability and raise alerts on data drift and performance decadence. Batter the model with explicit configuration files generate human review fallbacks and establish rollback steps such that the system is safe reliable and simple to run.

**Data Preprocessing**

**Understanding the Data**

The initial important step in developing valuable insights and predictive models is to understand the dataset. The data set, used\_device\_data.csv, has 15 attributes that describe the important information about smartphones and their resale value. There are categorical data, such as device\_brand and OS, and numerical data, such as screen\_size, rear\_camera\_mp, front\_camera\_mp, internal\_memory, RAM, battery, and weight. Alongside, temporal and usage-related attributes like release\_year and days\_used can be useful to add context to device age and lifecycle, whereas the response variables normalized\_used\_price and normalized\_new\_price, can be used to model continuous predictions and price classes. These attributes collectively comprise a complete list of the factors that affect the resale value of mobile phones.

A key element in interpreting the data is to determine the role of each attribute in modeling. Some of the features like screen\_size, internal\_memory and RAM are likely to affect the prices directly, as they determine the main performance and usability of a device. Likewise, rear\_camera\_mp and front\_camera\_mp are indicators of multimedia features, which are major consumer demand factors in the current market. Binary variables like 4g and 5g are technological compatibility and would most likely affect price levels depending on network preparedness. Conversely, factors such as device\_brand and OS reflect brand reputation and ecosystem preferences, which can have categorical effects on price that can interact with other technical specifications.

Lastly, data quality issues should be checked and transformation prepared. Missing values, inconsistent entries, or skewed distributions may be required to be fixed in order to have a robust modeling. As an example, days\_used and release\_year might have to be converted into derived attributes like “device age” in order to reflect the effects of temporal depreciation. Likewise, normalized\_new\_price can be used to provide a benchmark to contextualize resale value and improve model interpretability. Investigating the meaning, type, and distribution of each attribute in detail, we will be able to develop a solid base of exploratory analysis and predictive modeling, where business objectives are properly aligned with analytical results.

**Attributes Definition**

| **Variable Name** | **Description** | **Type** |
| --- | --- | --- |
| **device\_brand** | Brand or manufacturer of the smartphone (e.g., Apple, Samsung, Huawei, etc.) | Categorical |
| **OS** | Operating system of the device (e.g., Android, iOS, etc.) | Categorical |
| **screen\_size** | Screen size of the device in cm | Numerical (float) |
| **4g** | Whether the device supports 4G connectivity (1 = Yes, 0 = No) | Binary |
| **5g** | Whether the device supports 5G connectivity (1 = Yes, 0 = No) | Binary |
| **rear\_camera\_mp** | Rear camera resolution in megapixels | Numerical (int) |
| **front\_camera\_mp** | Front camera resolution in megapixels | Numerical (int) |
| **internal\_memory** | Internal storage capacity in GB | Numerical (int) |
| **ram** | Random Access Memory (RAM) in GB | Numerical (int) |
| **battery** | Battery capacity in mAh | Numerical (int) |
| **weight** | Device weight in grams | Numerical (int) |
| **release\_year** | The year the device was released | Numerical (int) |
| **days\_used** | Number of days the device has been used | Numerical (int) |
| **normalized\_used\_price** | Normalized resale price of the device (target variable for regression/classification) | Numerical (float) |
| **normalized\_new\_price** | Normalized new price of the device (benchmark to compare used price) | Numerical (float) |

**Summary of the Data**

The data set has 3,454 smartphone records and 15 attributes that represent the specifications of the device, usage patterns and prices. Categorical variables like device\_brand, os, 4g, and 5g give an idea of brand distribution, operating system distribution, and connectivity features. Continuous variables such as screen\_size have a large range of 5.08 to 30.71 cm, which indicates a large number of different types of devices, including small ones and some unusually large ones that may be outliers. Similarly, rear\_camera\_mp and front\_camera\_mp show a large range, with the maximums of 48MP and 32MP respectively, but some records have missing values (179 in rear camera and 2 in front camera). Memory related features like internal\_memory (0.01 GB to 1024 GB) and RAM (0.02 GB to 12 GB) also have a wide range with small numbers indicating data entry errors or extreme outliers. Battery sizes vary between 500 and 9720 mAh, so there is quite a variety of lifespan and performance of devices, and some entries are missing.

Other features emphasize the durability of the device and its costing. Weight ranges between 69g and the extreme 855g with 7 missing values that might need cleaning. The Release\_year is between 2013 and 2020 indicating that the data covers devices released within a span of eight years. The range of Days\_used is 91-1,094 days (or about three years), which coincides with the usual replacement cycle of smartphones. The ranges of the response variables, normalized\_used\_price and normalized\_new\_price are relatively stable, with used phone prices ranging between 1.54 and 6.62, with a median of ~4.4, and new phone prices between 2.90 and 7.85, with a median of ~5.2. These distributions show that most of the devices are within the expected ranges, but there are extreme outliers and missing data points that should be handled during the preprocessing stage to provide robust and reliable model performance.

**Structure of the Data**

The data is arranged in the form of a data frame with 3,454 observations and 15 variables, including both categorical and numerical characteristics of used smartphone specifications and prices. Categorical variables are device\_brand, OS, 4g and 5g, which are stored as character strings. These attributes include key descriptive features: the brand name (e.g. Honor, Samsung, Apple), the operating system (e.g. Android, iOS), and the support of 4G/5G networks. As these variables are now in character format, they probably will have to be converted to factors or dummy variables in order to use them in modeling, particularly in regression or classification exercises.

The numerical variables in the data set are technical specifications and physical characteristics of the smartphones. As an example, screen\_size is a continuous numeric variable that measures the size of the screen in cm, starting with compact models and going up to large-screen models. rear\_camera\_mp and front\_camera\_mp give the camera resolutions in megapixels, and internal\_memory and RAM give the storage and processing capacity, respectively. Other performance indicators are battery (mAh capacity) and weight (grams) that provide information on power sustainability and device design. These quantitative measures are the main predictors of the modeling of price, but they also need cleaning because of missing values and possible outliers.

Time and price related features provide additional complexity to the data. release\_year and days\_used will be important in depreciation analysis since it is essential to understand how old a device is and how much it has been used. Lastly, the two dependent variables normalized\_used\_price and normalized\_new\_price are stored as numeric data, where the former is the main target variable to be predicted or classified. The availability of both normalized prices of used and new cars makes it possible to compare and contrast, in a bid to reveal how usage and specifications affect depreciation. On the whole, the nature of this dataset offers an adequate balance between categorical and numerical data that can be used to conduct exploratory analysis and predictive modeling that are consistent with the objectives of the project after preprocessing.

**Numerical Data**

The data set has a number of numerical variables that indicate the technical specifications, physical attributes, and lifecycle factors of smartphones. Core performance-based attributes are screen size (cm), rear camera (mp), front camera (mp), internal memory (GB), and RAM (GB). These parameters are directly related to hardware capabilities, and the higher the value, the more likely it is that the device is premium, and, accordingly, it can be sold at a higher price. Similarly, battery capacity (mAh) and weight (grams) provide additional details about the usability and the quality of the device, but atypically high or low values can be an indicator of an outlier. Temporal dimensions are captured in the form of release\_year, which indicates the technological generation, and days\_used, which indicates wear and depreciation. Collectively, these continuous variables offer a detailed numerical basis of explaining and predicting price variability.

The two variables that define the core modeling objectives are the target outcomes normalized\_used\_price and normalized\_new\_price which are numeric variables. The normalized used price is the target variable in regression and classification tasks, and the normalized new price is used as a reference, which enables the study of depreciation trends across brands, features, and time of use. Preliminary summaries indicate that a number of attributes have wide ranges (e.g., internal memory up to 1024 GB, RAM up to 12 GB, and battery capacity up to 9720 mAh), which indicates the presence of extreme values that require special treatment during preprocessing. In general, the numerical data does not only reflect the necessary device features but also provides a chance to identify trends, relations, and predictors that have a significant influence on the resale value of smartphones.

**Categorical Data**

The data set contains a number of categorical variables that describe non-numeric but very powerful features of smartphones. These are device\_brand and OS, 4g and 5g which are all stored as character strings. Device\_brand is the manufacturer, which usually has significant brand equity implications on prices, such as with Apple devices, which tend to have higher resale values than less well-known brands. OS refers to the operating system used by the phone, whether Android, iOS, or other operating system, which represents ecosystem preference and user loyalty. The binary variables 4g and 5g indicate whether the device is 4G or 5G compatible, and newer technologies such as 5G will probably attract higher prices. These variables are categorical and thus will have to be encoded (e.g., dummy variables or factors) to be used in statistical modeling, but their inclusion is essential since they capture consumer perceptions, technological adoption, and brand-driven value that cannot be explained by numerical specifications alone.

**Missing Values**

**A graph of a device

AI-generated content may be incorrect.**

The analysis of the data set shows that there are 202 missing values distributed among some of the numerical features, whereas categorical features are complete. The most impacted attribute is rear\_camera\_mp with 179 missing entries, which is approximately 5.18 % of the dataset. Other variables like front\_camera\_mp, internal\_memory, RAM, battery, weight also have small percentages of missing values, which are 0.05 to 0.20. These percentages, although relatively small in comparison with the total number of observations, 3,454, may cause biased results, low model accuracy, and difficulties in exploratory data analysis, should they be left untreated.

In this project, the strategy to be used is deleting rows containing missing values (na.omit). This strategy is legitimate since the percentage of missing data is not high, with the maximum of 5.18 % of rear\_camera\_mp. Deleting incomplete cases eliminates inconsistency in the dataset, the possibility of injecting artificial values by imputation and preserving data integrity. The deletion of 202 records is unlikely to substantially decrease the statistical power or the generalizability of the models since the dataset is large enough.

This strategy will make the data set complete and without any missing values, which will enable the transition to the following data preprocessing and model construction stages. More sophisticated methods like imputation may be used but they may add noise or assumptions that are not supported by the data. The removal of the missing rows will result in a clean and reliable dataset, which will form a solid basis of exploratory data analysis, feature selection, and predictive modeling, which are all important to the accurate and business-relevant results of used smartphone price analytics.

**Count of Zeros**

The examination of zero values in the data indicates that there are 39 instances of which all are in the front\_camera\_mp attribute. This is approximately 1.13 % of the data, which means that few devices are reported to have no front-facing camera. In terms of data quality, these values can be true in instances where older or low-end smartphones do not have front cameras at all. But they may also be placeholders to missing information, particularly when the devices are new and should have such a capability. It is important to identify whether these zeros are real specifications or data entry anomalies before modeling.

Because the occurrence is confined to one feature and to a small proportion of the data, the options available include leaving the zeros unchanged in case they represent realistic values, or treating them as missing data in case they represent errors or inconsistencies. When treated as missing, the same approach employed above (removal of incomplete rows) could be used to maintain data integrity. Alternatively, one may also consider imputation methods like using median or mode value of front\_camera\_mp but this might be biased. Finally, for the dataset to be consistent and dependable in predictive modelling, the handling strategy needs to be able to balance technical accuracy with business context i.e., Only rows with obviously impossible entries should be dropped carefully.

**Visualizing the Variables**

**Device Brand Distribution**

A graph of a number of blue bars

Description automatically generated with medium confidence

The brand distribution of devices indicates a very heterogeneous market, with some brands such as Huawei, Samsung, Oppo, and LG being the most common, and many other minor brands providing a smaller number of devices. In particular, the number of devices in the Others category is the highest, about 500, which indicates that a large number of less popular or niche brands are observed in this category. This spread indicates that although the top brands have a huge presence in the resale market, there are still a large number of smaller players, which signifies a competitive and fragmented used smartphone market.

**Operating System Distribution**

A green bar graph with black text

Description automatically generated

The data distribution of operating systems is strongly skewed toward Android, which comprises most of the devices. The data represented by iOS, Windows and Others is very minimal. This is reflective of the global trends of smartphone markets where Android dominates the majority of the market share. Nevertheless, the number of iOS devices is lower, but they are usually more valuable in resale, which can also affect pricing models and indicate OS as an essential categorical predictor.

**4G Support**

The 4G support distribution indicates that the majority of smartphones in the dataset support 4G, whereas a smaller number does not. This is in line with the release years (2013-2020), a time when 4G technology became the norm on mid-range and flagship phones. As 4G has a considerable impact on usability and connectivity, this aspect is likely to have a positive impact on resale value, whereas non-4G devices will be in lower price brackets.

A screenshot of a computer screen

Description automatically generated

**5G Support**

Unlike 4G, most of the devices are not compatible with 5G and only very few smartphones have the capability. This observation is consistent with the fact that the adoption of 5G started in the later years of the dataset (2019-2020). As 5G gains mainstream popularity, these handfuls of 5G-enabled devices will probably be marketed as premium products in the resale market, and will fetch higher prices than their predecessors that are 4G-only.

A purple square with a number of text

Description automatically generated with medium confidence

**Screen Size Distribution**

A graph of a screen size distribution

AI-generated content may be incorrect.

The distribution of the screen size shows that the majority of smartphones are between 12 and 15 cm with a maximum at 13 cm which is the typical size of smartphones. There are however some extreme values that are greater than 20 cm that could be tablets or data entry errors. The tendency to cluster around the middle screen sizes indicates that consumers prefer the standard size of smartphones, whereas the existence of outliers demonstrates the necessity to clean the data before modeling, to obtain proper insights.

**Rear Camera MP Distribution**

A graph of a camera distribution

Description automatically generated

The graph of the rear camera megapixels indicates that the majority of devices are between 5 MP and 13 MP, with a slight peak at 13 MP. Very few devices have extremely high camera resolutions with outliers going up to 48 MP. This implies that although the dataset has higher megapixel values, most smartphones represent mid-range camera capabilities which were the standard in the market between 2013 and 2020 release years.

**Front Camera MP Distribution**

A graph of a camera distribution

Description automatically generated

The front camera megapixels is right-skewed with many devices having 0-5 megapixels, indicating the early smartphones or entry-level devices with basic selfie cameras. Another minor peak is around 15 MP, which is more recent mid-to-high-end smartphones. Some of the outliers go over 20 MP to as high as 32 MP, probably high-end flagship phones. This distribution shows that the quality of front cameras has changed with time considerably.

**Internal Memory Distribution**

Internal memory distribution is also highly skewed with the majority of devices falling to within 16 GB, 32 GB, and 64 GB storage capacities, which were typical storage sizes at the time of the dataset. Some of them have significantly larger storage, up to 1024 GB, but they are exceptional cases that probably represent high-end flagship or special-edition devices. This broad range demonstrates the significance of internal memory as a robust pricing differentiator, with bigger capacities typically corresponding to higher resale prices.

A graph showing the internal memory distribution

Description automatically generated

**RAM Distribution**

The distribution of RAM indicates that 4 GB is the most popular configuration by far, with the rest of the devices being concentrated around it. Fewer devices have 2 GB or 3 GB, which is characteristic of older or low-end smartphones, and a few have larger values up to 12 GB, which is characteristic of more advanced devices. The obvious prevalence of 4 GB indicates its position as a standard in the years of the dataset, but the higher RAM values should have a positive effect on the resale prices.

A graph with purple rectangular bars

Description automatically generated

**Battery Capacity Distribution**

A blue graph with black text

Description automatically generated

Battery capacities are distributed broadly with most devices between 2000 mAh and 4000 mAh, which is the industry standard in the mid-2010s. The right tail is long with values going beyond 7000 mAh to 9720 mAh which are probably tablets or outliers. The shape of the overall suggests that most smartphones have mid-range batteries, but extreme values should be handled with care during preprocessing, as they may distort the modeling results.

**Weight Distribution**

A graph of weight distribution

Description automatically generated

The distribution of weights indicates that the majority of smartphones are within the range of 120 grams and 200 grams, which is the normal weight range of the usual smartphones. Some of the devices weigh far more than 400 grams, with outliers over 800 grams, which are probably tablets or erroneous records. Extreme values are uncommon yet prominent and therefore, during preprocessing, these outliers must be evaluated with care since they can have a disproportionate impact on model outcomes when not addressed.

**Release Year Distribution**

A graph of a number of years

Description automatically generated

The distribution of the release year is between 2013 and 2020, with the largest number of devices concentrated in the range of 2013-2015. The devices per year then decline a little, with a lesser peak in 2019. This pattern can be attributed to the bias in the collection of the datasets since older models are more likely to be sold in secondary markets than the latest releases. It shows the significance of the age of devices in depreciation and pricing of used smartphones.

**Days Used Distribution**

The distribution of days used indicates a fairly even distribution between 100 and 1,100 days with a peak of 600-700 days. This is an indication that a large number of smartphones in the dataset are resold after approximately 1.5 to 2 years of use, which is consistent with upgrade cycles. The fact that there are devices with shorter and longer lifespan however, suggests that there is variability in how users behave, and this aspect is likely to be a critical factor in price prediction models.

A graph of a number of days

Description automatically generated

**Normalized Used Price Distribution**

The distribution of the normalized used price is roughly bell-shaped, with the majority of the devices priced between 4.0 and 5.0 on the normalized scale. The number of devices that are less than 3 or more than 6 are very small, which means that most of the resale values are concentrated within a central range. This stable distribution is suitable to support regression modeling because it implies that the response variable is well-behaved and that there is limited skewness, although there may still be outliers on both sides of the distribution.

A green bar graph with black text

Description automatically generated

**Normalized New Price Distribution**

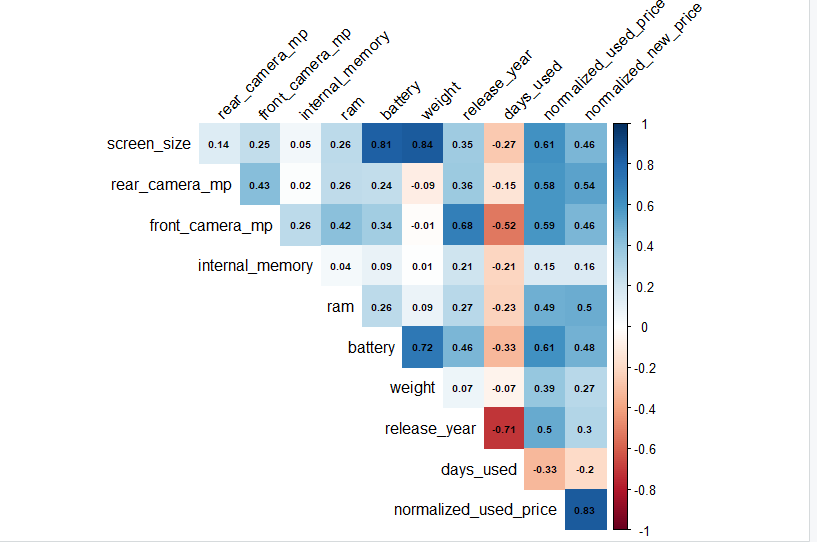
A blue graph with black text

Description automatically generated

The distribution of the normalized new price is also bell shaped with the majority of the values between 4.5 and 6.0 with a mean of 5.0. The spread is a bit wider than that of the used price distribution and the higher-priced devices reach nearly 8.0. The trend affirms that resale values are also compressed compared to the new prices due to depreciation effects. Incorporating used and new normalized prices into the analysis will assist in capturing the effect of certain features on value retention over time.

**Predictor analysis and relevancy**

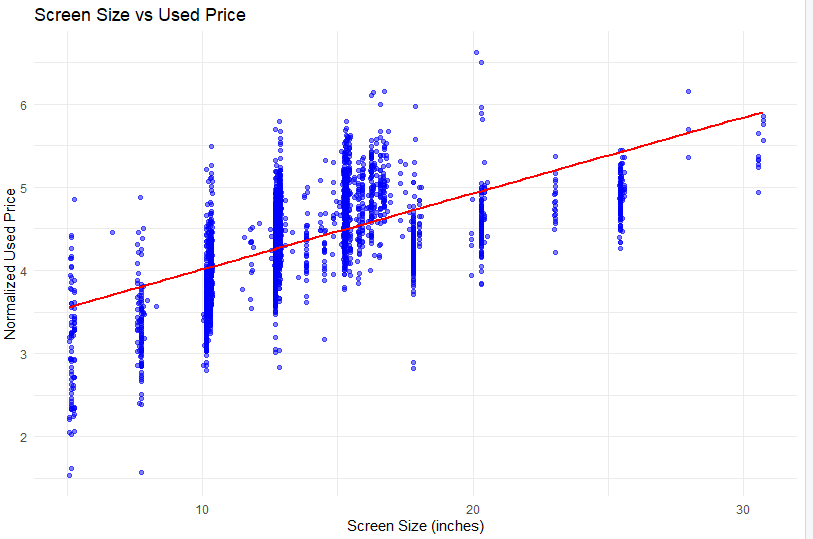
**Correlation Plot**



The heatmap of correlation shows some business relevant and intuitive relationships. Normalized used price is most strongly driven by normalized new price (= 0.83), which affirms that the resale value is pegged to the original market value of the device. In addition to this anchor, technical specifications indicate moderate positive relations with used price: larger screen\_size and battery (approximately 0.6) as well as bigger front and rear camera megapixels (around 0.59 and 0.54) and RAM (around 0.49) are all related to high resale prices. Newer release year is also positively related to used price (≈0.50) and days used negatively (≈ −0.33) which includes depreciation due to wear and age. There are prominent internal relationships outside the target: screen size vs. battery is very high (=0.8+) and battery vs. weight very strong (=0.7), and the release year vs. front camera mp is strong ( =0.68) because of co-evolution of camera technology and newer models. Similarly, release year vs. days used is significantly negative (= -0.71), because newer phones have been in use less.

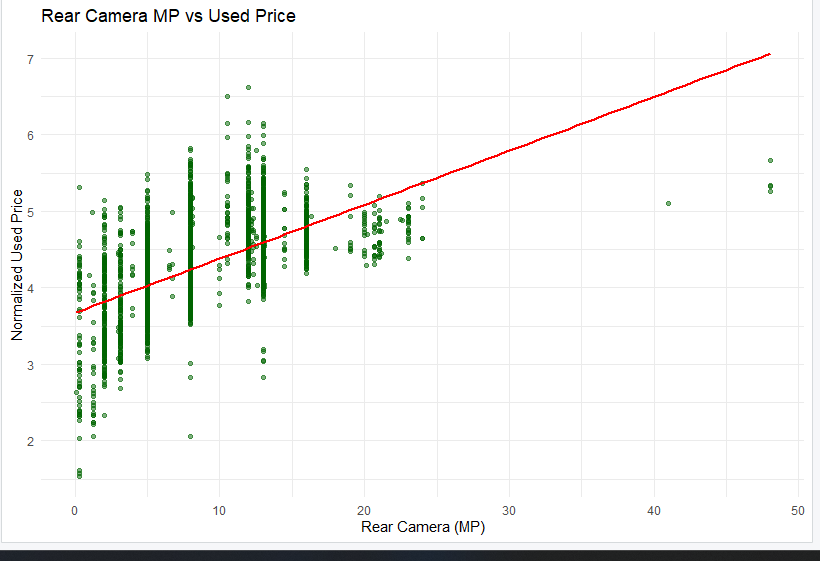
On the modeling front, these clusters indicate the possibility of multicollinearity between size/power/weight (screen\_size, battery, weight) and between generation characteristics (release\_year with camera specs and days\_used). In the case of linear models, think about VIF tests, and, in practice, regularization (Ridge/Lasso/Elastic Net) or pruning/merging highly correlated predictors (e.g. keep battery, drop weight, or compute device age = current year - release year). Since normalized new price is the largest contributor to signal, determine whether your business question is absolute resale prediction (keep it) or value retention; with the latter, use ratio of model used/new price to decrease target-predictor coupling. Lastly, keep in mind that correlations are linear summaries: the scatterplots that you are generating (with trend lines) are necessary in order to identify non-linearities, heteroskedasticity and outliers that a heatmap cannot identify on its own.

**Screen Size vs Used Price**



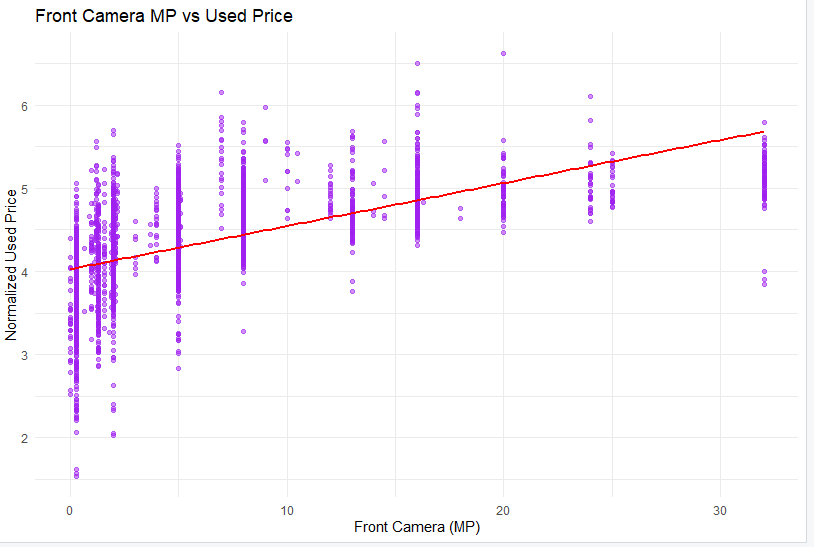
It can be observed in the scatterplot that larger screen sizes tend to fetch a higher resale price, and it is clear that the regression line has an upward trend. This aligns with the tendencies of consumers to purchase larger displays that increase media consumption, productivity, and gaming. On the business side, this understanding can be used by resellers to focus on screen-size in their marketing or pricing strategies, so that larger sizes will sell at a higher price. But extreme outliers (very small or very large screens) are to be scrutinized, as these may indicate niche devices or data entry anomalies that can bias predictive accuracy.

**Rear Camera vs Used Price**



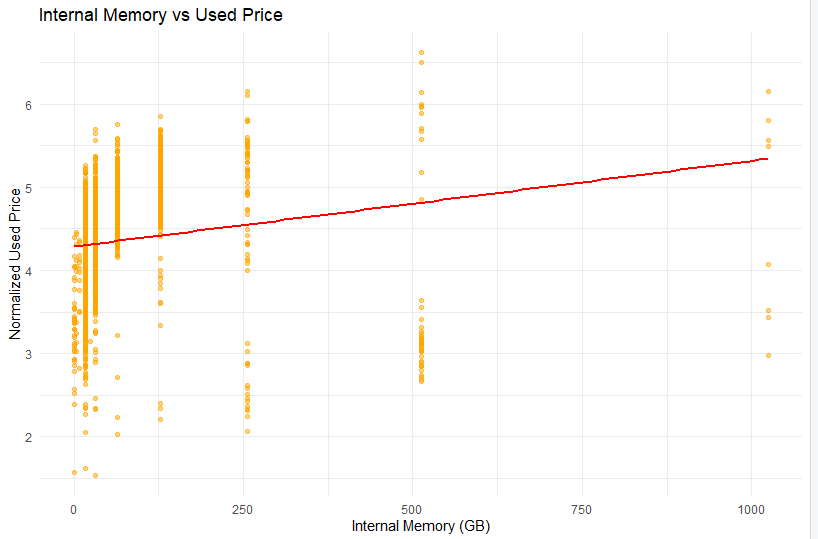
The used price has a positive correlation with the rear camera resolution, because consumers tend to want smartphones with higher photography capabilities. The trend indicates that the more megapixels the device contains, the more value it gets in the resale market. With regards to business intentions, this aspect enables the use of camera features as a point of distinction in justification of the increased prices, and in the context of problems such as underestimation, it would be a clear sign on how fairly the resold products should be priced. Very high-MP outliers need handling with care because they may be of premium models or inflated specifications that do not necessarily correlate with the resale value.

**Front Camera vs Used Price**



Front camera megapixels also have a positive but steady relationship with resale value, but the association seems to be a little weaker than with rear cameras. This is logical in the current market where the culture of selfies and video calling is one of the primary motivation factors in user preferences. Business wise, resellers must consider quality of front camera as a pricing factor, especially among the younger age groups or those who are content creators and value this aspect. Here the business issue of price transparency is enhanced, since the quality of the camera is an easily observable and justifiable metric by which a customer can assess whether a device is worth its quoted price.

**Internal Memory vs Used Price**



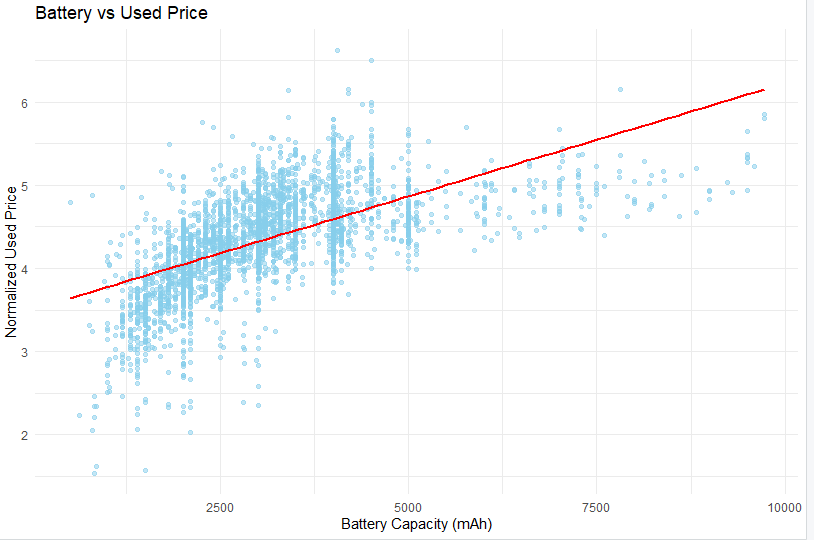
The used price shows a weak positive correlation with internal storage capacity but the scatter is wider and concentrated around the typical levels (16, 32, 64, 128 GB). This implies that the resale value generally increases with a larger storage capacity, but the marginal benefits become zero at a higher storage capacity, except with high-end equipment. This helps in developing storage-based pricing levels, where there are fewer chances of creating pricing below the capacity needed by business goals. One such aspect of the business problem that needs to be taken into account is verifying extreme values (e.g. 500 GB or 1024 GB), unless they will misprice or falsify a model without validating and cleaning them.

**RAM vs Used Price**



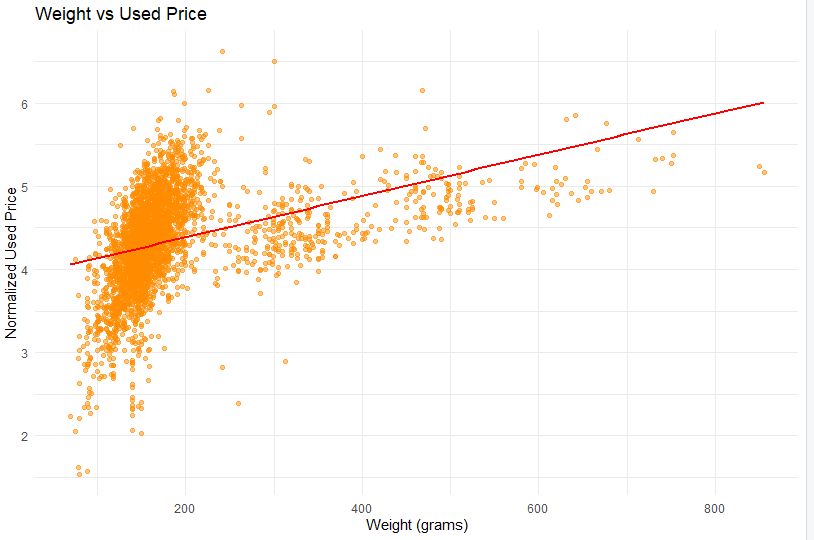
As a performance indicator, RAM is strongly positively correlated with resale price. The more RAM in the device, the faster and more future-proof it looks, so it is more desirable in the secondary market. Regarding business purposes, RAM features a well-defined and well-explaining characteristic that facilitates competitive prices and justifies high-quality devices premiums. To solve business issues, this relation assists in removing uncertainties in the pricing strategies and the customers are aware of the reason the devices with high RAM have a high price, but the multicollinearity between the release year and the new price must be taken into consideration in predictive modeling.

**Battery vs Used Price**



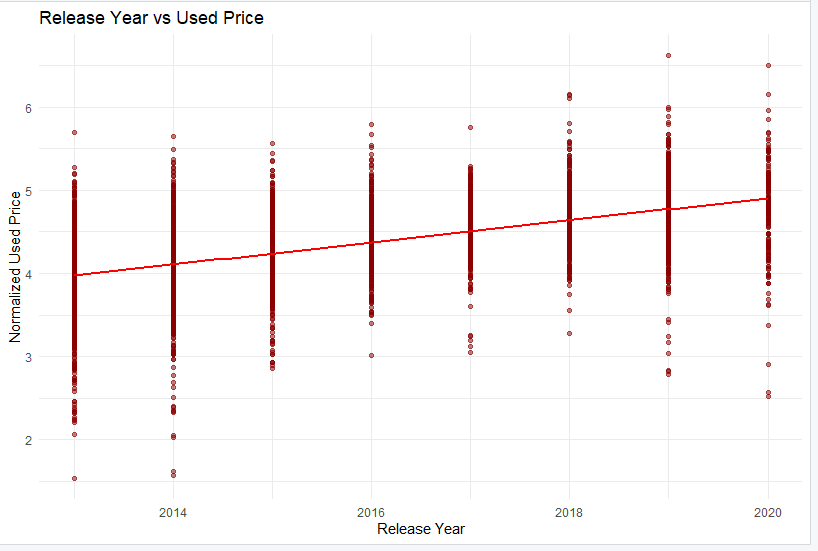
Battery capacity has a very high positive correlation with resale value as well, indicating that consumers like longer-lasting devices. They are perceived as the more dependable phones, particularly when used heavily, and thus the big battery phones attract the second-hand market. This is important to the business because resellers can comfortably charge more when selling devices with batteries above 4000 mAh. Nevertheless the business issue of fairness in pricing needs to consider outliers such as very large battery values (7000-9000 mAh) which might be part of tablets, or specialty models that can bias predictive models unless treated carefully.

**Weight vs Used Price**



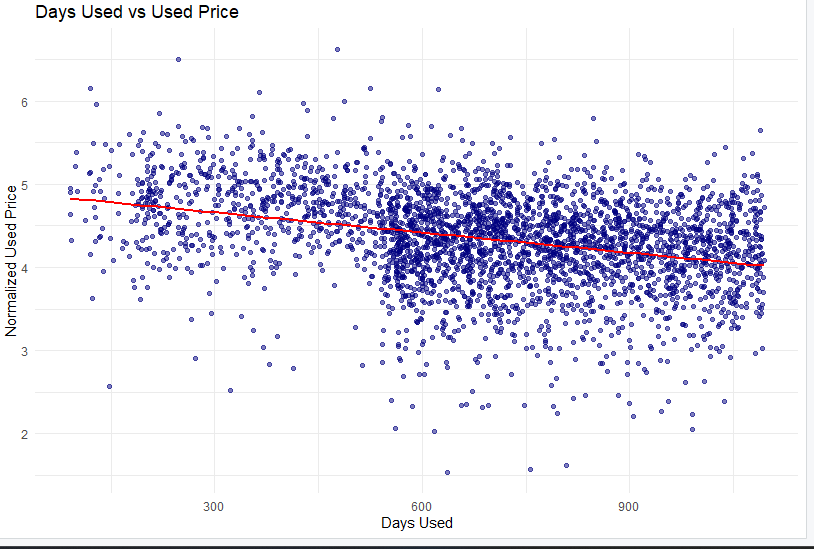
The plot exhibits a positive and weak association between weight and resale price, but most of the data is concentrated at standard smartphone weights (100-200 grams). Devices that are larger, like in size and weight, or quality of construction, could warrant a slight increase in resale values. Weight is not an independent driver as far as business objectives are concerned, it must be taken together with screen size and battery capacity. Viewing this as a business problem, the outliers above 400 grams can be extreme, or can reflect an error in the data, and need cleaning to avoid bias in pricing forecasting.

**Release Year vs Used Price**



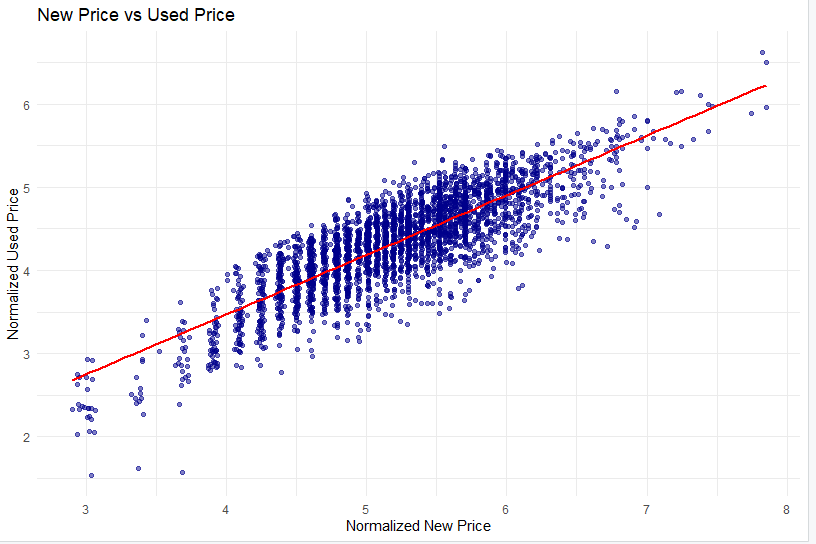
Unsurprisingly, the year of release is linked to stronger resale values as newer release years, which confirms the idea that technological recency is a major force behind market prices. This can be used to advance business interest as resellers can include the age of devices as a distinct pricing tier variable that allows newer devices to be priced as higher quality. It addresses the problem of depreciation in business-related questions by estimating quantitatively the rate at which older models decline in value so that the practice of clearly and consistently marking down products can be adopted to earn consumer trust without jeopardizing the business.

**Days Used vs Used Price**



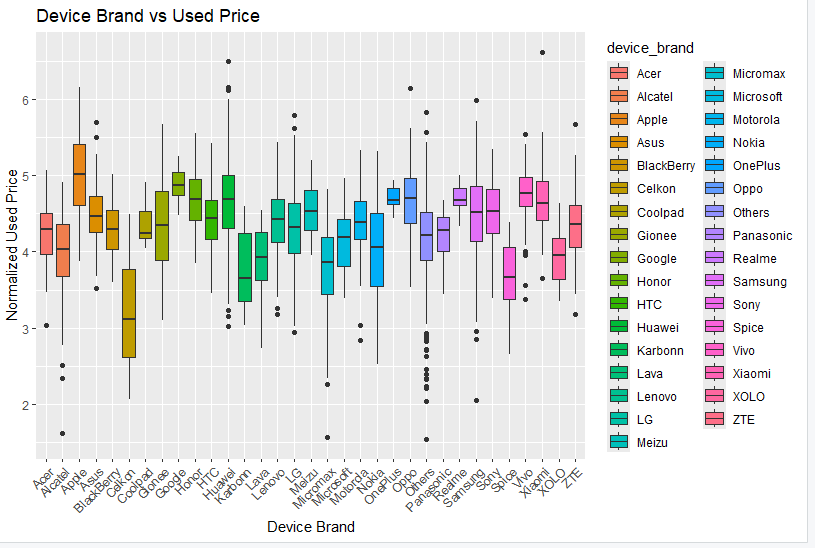
The scatterplot indicates that there is a definite negative direction in that the higher the number of days of use, the lower the resale prices. This is in line with the business knowledge that wear and tear decrease the perceived value, despite the same specifications. This knowledge can help the resellers to dynamically price their products according to the time in which they will be used, giving buyers transparency and avoiding under-valuing of under-utilized devices. Associated with business matters, this aspect helps in standardization of principles of depreciation to prevent any confusion and to offer a coherent resale policy irrespective of the condition of the devices at question.

**New Price vs Used Price**



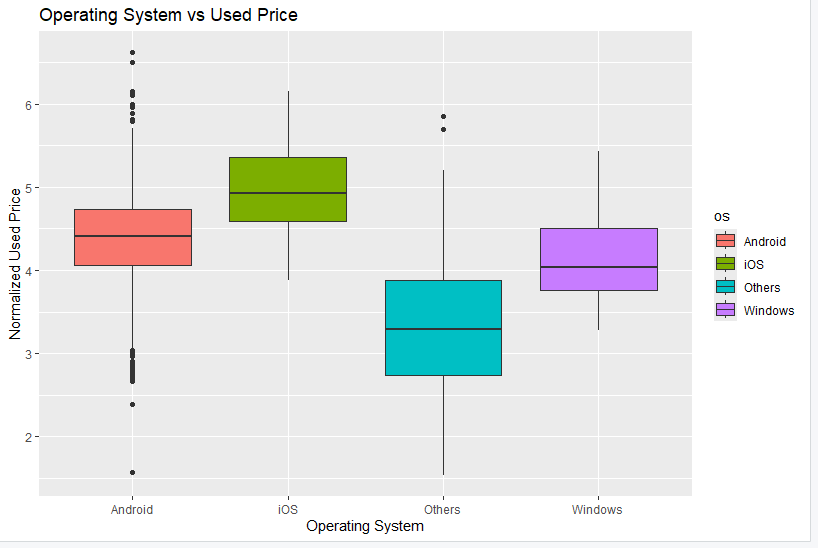
The strongest relation is that between new and used prices where there is a high positive correlation. This ascertains that the original retail value is the main anchor when it comes to the determination of resale prices. The variable plays a core role in business objectives in order to attain appropriate forecasting and fair market adaptation as the customer comparatively instinctively examines the original price. The business issue in this case is dependency: on the one hand, it offers a great predictive power, but, on the other hand, overdependence on new price disregards other value-retaining points, such as brand, RAM, or battery, thus, in modeling, it should be combined with technical specifications.

**Device Brand vs Used Price**



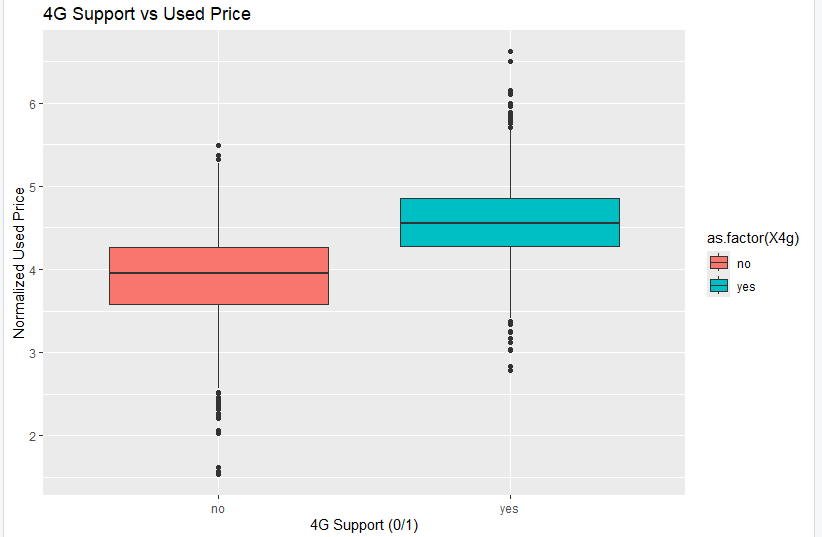
It can be seen that brand is one of the most effective categorical values that drive resale value, with Apple, Samsung and some premium brands always sticking with higher prices, than relatively unknown/generic brands, as shown by the boxplot. This helps to achieve business objectives by showing how the reputation of a brand has a direct effect on consumer readiness to pay it, which is a crucial parameter in price formulas. This insight can be used to solve business problems by avoiding undervaluation of superior brands with reasonable expectations of low-equity brands to enable fair and competitive pricing in a fragmented market.

**Operating System vs Used Price**



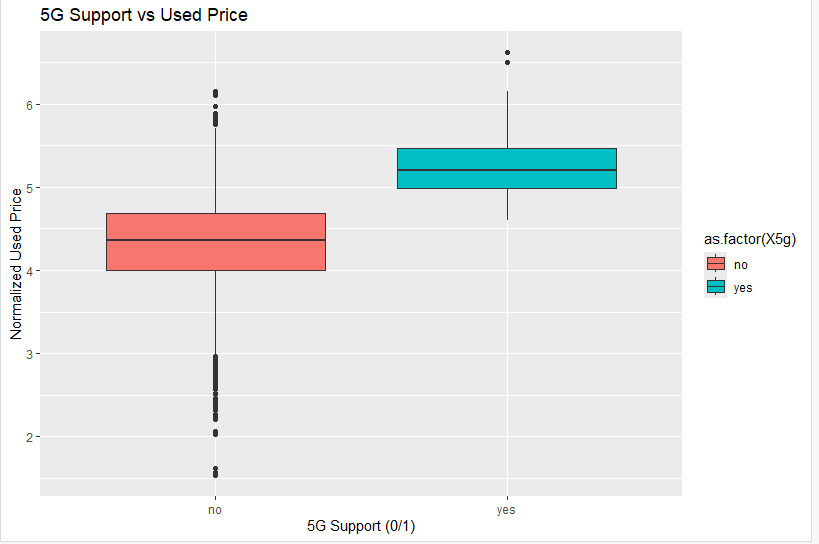
The operating system also affects the resale prices, and all iOS devices are obviously priced at higher values in resale than Android, windows, or other operating systems. This indicates ecosystem stability perceptions of consumers, device life perceptions, and brand equity perceptions about operating systems. Regarding business objectives, this variable reinforces pricing models by capturing ecosystem-driven value, and regarding business problems, it can overcome the price transparency challenge, where buyers are able to understand why iOS devices can sell at higher prices, and sellers can explain why they can justify higher prices when based on OS advantages.

**4G vs Used Price**



The average resale value of 4G devices is significantly higher than the average resale value of non-4G devices, indicating the role of connectivity in the decision-making process of buyers. Since 4G became a standard within the period covered by the dataset, this characteristic means that older devices are at an apparent disadvantage when reselling in the market. This variable is helpful in terms of business since old and new models could be sold in a differentiated way, with sellers not overpricing non-4G devices and buyers not feeling deceived by their buying choices. It is also the solution to the issue of being undervalued as the premium in relation to connectivity standards is quantified explicitly.

**5G vs Used Price**



The boxplot shows that 5G support is linked to a significant resale premium and consumer demand to have next-generation connectivity, even though the adoption window was relatively brief in the dataset. In the context of business objectives, this makes it essential to ensure that predictive models have cutting-edge technological features, which have a profound impact on value retention. In business problem terms, it points out the risk of underpricing early 5G equipment - resellers who do not factor in the cost of this premium might miss sales, and buyers may have a clear understanding that 5G support is worth a higher asking price.

**Dimension reduction**

The common use of dimension reduction in a large number of predictors (and many are likely to be redundant or very correlated) can result in multicollinearity and a loss of interpretability in the model. But, in this project, the data set has a reasonable amount of attributes (15 variables, some categorical and some numerical), which is not overwhelming and does not impose a curse of dimensionality. Every piece of data in the data set is an important business or technical quality of a smart phone, like screen size, RAM, megapixels of the camera, or brand that all give different clues about resale value. The dataset is not too huge in either dimensionality or size so dimensionality reduction could cost it critical predictors that may only be needed when making business-related interpretations.

The other reason why dimension reduction is unnecessary in this case is because the predictors can be interpreted in a direct business sense. Attributes such as battery capacity, year of release, or 5G capability are self-explanatory to both sellers and buyers and contribute useful value to the resale pricing model. Had we applied such techniques of dimension reduction as PCA (Principal Component Analysis) or factor analysis, the obtained components would be abstract combinations of variables that are not easy to explain to stakeholders in the business. As an example, a key feature that would integrate screen size, weight and battery capacity would not offer the same level of clarity in decision making as would operationalizing these aspects independently, particularly given that customers and resellers assign importance to these attributes independently.

As indicated by the correlation analysis conducted above, there are moderate correlations between some of the predictors, including the screen size and the battery or release year and camera characteristics. The correlations, however, are not near to levels of severe multicollinearity that would destabilize regression or classification models. The current machine learning algorithms, including random forests and gradient boosting, are resistant to correlated predictors and can tolerate redundancies internally, without the need to reduce dimensions. By keeping the original variables therefore, the model does not lose any relevant information whereas it is free to give weight to the significance of every predictor.

Lastly, regarding business, interpretability is equally significant just like predictive accuracy. Reduction of the dimensions would make it more difficult to communicate results to resellers and customers since clear, understandable attributes would be converted into latent dimensions. As clarity of predictors is crucial to the achievement of one of the main purposes of this project, ensuring clarity, transparency will be maintained by the retention of all predictors. Thus, in relation to the numerical size of the manageable dataset, the interpretability requirements, and the moderate degree of correlation, the dimension reduction is unnecessary in this analysis.

**Data transformation**

Data transformation was used mainly in this project as the first step in preparing the data to be analyzed and modeled properly by treating missing values, zero values and inconsistencies. The first step was to address missing data, where approximately 202 records had not been fully completed, mostly the features of rear\_camera\_mp and few others such as front\_camera\_mp, RAM and weight. Given that the percentage of missing values was not very large (at most about 5%). Given the low rate of missingness, we assume values are missing completely at random (MCAR), making listwise deletion acceptable. This approach did not add any artificial bias to the dataset. Likewise, zero values, notably in front\_camera\_mp attribute, were evaluated with great care; although some zeroes represented real older models with no front camera, some were possible placeholders. These values were audited and stored where they made sense and the transformations did not alter the business meaning of the data.

The response variables normalized\_used\_price and normalized\_new\_price were already normalized, and pre-scaled so that comparisons can be made across different devices, and these variables are not biased by absolute currency values. In addition, the categorical variables like device brand, operating system, 4g, and 5g were coded as factors to prepare them to be used in the statistical modeling and visualization. This conversion allowed easier analysis in boxplots and (eventually) in predictive models using dummy variables or one-hot encoding where necessary. Quantitative predictors such as screen size, RAM and battery were left in their raw units because their scale has business sense and is readily understandable by both sellers and buyers. In general, these changes assured clean, consistent, and directly interpretable dataset, and retained all significant business indicators in resale price analysis.

**Data partitioning methods**

The data partition is one of the most crucial processes in the predictive model because it ensures that the performance of a model is objectively and unbiasedly determined. Given that a model could be trained and tested using the same data, the model can merely memorize the patterns and not learn to generalize hence overfitting. By dividing the data into distinct sets, it is possible to train the model on one set and then test or evaluate it on another, which mimics real conditions in which the model is used to process unseen data. This is done to build confidence in the strength and dependability of the model in making business decisions, especially in competitive markets such as used smartphone pricing.

There are a number of methods that can be employed to partition the data based on the analytical objective. The simplest way of dividing the data is the train-test split where the data is separated into two, usually 70-80% training and 20-30% test. A stricter method is known as k-fold cross-validation, in which a dataset is divided into k equal parts and a model is trained and validated k times, with a different part being used each time. This minimizes the difference in the performance measures and makes the estimates more reliable. Another approach that provides the assurance that significant groups (e.g., brands or OS) are proportionately represented in the training and testing sets is stratified sampling, which is particularly meaningful in case of the unbalanced dataset.

Data partitions are mainly utilized to measure the objective performance of the model and to optimize hyperparameters prior to deployment. The model is able to learn predictor-target relationships with the help of the training set and the unbiased predictive accuracy is given when using the test set. Some of these might also include a third validation set, particularly in deep learning or complex models, to tune the parameters without visiting the test set. Partitioning is one method of assuring business applications like used smartphone price analytics that the predictive models are not merely performing well on historical data, but in fact will perform when applied to new equipment, and that sellers and buyers will make competitive and fair prices.

**Model selection**

**Linear Regression**

The linear regression is one of the most simplistic methods of prediction that approximates the resale price of smartphones as a linear sum of the predictors of screen size, RAM, battery, and camera megapixels. It is very interpretable and gives direct coefficients that indicate the extent to which each feature adds to price and as such is appealing to business transparency. It however presumes linear relationships and is prone to multicollinearity among predictors, that is, it does not necessarily reflect complex relationships in the data.

**Stepwise Regression**

The stepwise regression is an extension of the linear regression, which automatically selects the most relevant predictors using an iterative procedure of adding or removing them on the basis of statistical measures, such as AIC or p-values. In the case of smartphone pricing, this can assist smartphone manufacturers in the process of streamlining the model, in that it seeks to only highlight the most important features and excludes the unnecessary ones, thereby enhancing interpretability and decreasing the risk of overfitting. The trade-off is that in selecting, when predictors are strongly correlated, this method can be unstable, and it can fail to point to nonlinear effects.

**Random Forest Regression**

Random forest regression is one of the ensemble techniques that construct a number of decision trees and averages their results to estimate smartphone resale values. It is resistant to multicollinearity and is able to describe complex nonlinear interactions among attributes like storage, brand and connectivity. In contrast to linear models, it automatically deals with interactions, and it is not so sensitive to outliers. Its disadvantage is less interpretability because it becomes more difficult to explain precise feature contributions, but feature importances scores can also be useful.

**Logistic Regression**

Logistic regression is applied when the objective is classification, that is, when one wants to group smartphones into high and low tiers of resale price. It approximates the likelihood of a device falling into a category based on the predictors such as release year, 5G support or RAM, which are modeled with a logistic function. It is simple, interpretable, and powerful in the event that relationships are roughly linear in the log-odds space. Nonlinearities or multicomponent feature interactions are, however, not well represented in it as they are in tree-based classifiers.

**Decision Tree classifier**

A decision tree classifier segments the data into groups depending on feature values, i.e. whether the device is 5G or has more than a specific amount of RAM, to determine price categories. It has a well defined rule based framework that is most interpretable and intuitive to both technical and business stakeholders. Decision trees are well able to capture nonlinear relationships and interactions but are vulnerable to overfitting when not pruned or regularized appropriately, particularly when using many features.

**Random Forest Classifier**

The random forest classifier mitigates the drawback of using a single decision tree by using a huge number of trees constructed on bootstrapped datasets and voting together. It greatly increases the precision of smartphone pricing levels, minimizes overfitting, and generates consistent forecasts despite the existence of noise or correlated predictors. In comparison with a single tree, it is more predictable because it provides rankings of the importance of the features that have the most effect on price categories, without sacrificing predictive ability to business use.

**Modeling and Implementation**

**Linear Regression**

The main business objective of the project is to make it possible to transparently and correctly price used smartphones. Linear regression is a direct way of achieving this aim because it offers a statistical tool that correlates the smartphone features with the resale value including screen size, RAM, battery capacity, and camera quality. This model will turn the abstract objective of predicting fair and competitive prices into a quantifiable framework to which each coefficient is interpreted to determine the extent to which a unit change in a feature affects the normalized used price. This simplifies the process of resellers and customers to perceive the roles of technical specifications and brand features in the process of arriving at prices to meet the interpretability and trust requirements of the secondary market.

The cleaned dataset was trained and assessed on the 80/20 train-test split using a linear regression model. The model had a Test R2 of 0.849 and Test RMSE of 0.235; this implies that it forecasts almost 85% of used smartphone prices and the prediction error on unknown data is relatively small. These findings emphasize the point that the model can be useful in describing the key price determinants, including device performance, age and brand without losing predictive power and simplicity. The shape of the price distribution is also bell-shaped, the reason why the use of linear regression is suitable is because the response variable fits the conditions of the tool.

There are two implications of the application of linear regression- first, resellers can use a clear formula to justify the changes in price, and second, customers will understand a visible correlation between the characteristics and the pricing. Nonlinear interactions of the smartphone specifications and that, multicollinearity exists between the predictors (e.g., screen size and battery), could therefore be simplified by the linear regression. Although the model is performing well and it is a well-developed baseline, the more complex models such as the Random Forest or Gradient Boosted Trees may be used to obtain additional improvements in the accuracy by recognizing complex patterns. However, linear regression is a very interpretable and business friendly option to implement at the beginning.

**Stepwise Regression**

The objective of the project is to estimate the resale values of used smartphones in a manner that is accurate, simple, and able to be interpreted. Stepwise regression helps to achieve this objective, since it automatically chooses the most important predictors of price and excludes redundant or less significant variables. This is in line with the business requirement to pay more attention to the factors that have the most impact, which include, RAM, battery capacity, year of release and brand name, hence resellers and buyers can understand well why some specifications are more effective in price determination than others. Stepwise regression removes unneeded predictors to produce a model that is lean and easily understandable and applicable to pricing decisions.

A stepwise regression model fitted to the cleaned data had a Test R 2 of 0.846, and a Test RMSE of 0.238. This result is quite similar to the performance of the full linear regression model, so that a large part of the predictive power can be preserved even with the less relevant variables eliminated. The very slight loss in explanatory power (between R squared = 0.849 to 0.846) is compensated by the fact that the model is less complex and can be interpreted more readily. This shows that stepwise regression is useful in simplifying the model without compromising on the predictive strength of the model, hence is applicable in business scenarios where clarity of the feature contribution is cherished.

The most important implication of stepwise regression is that in addition to having almost the same level of predictive accuracy as an entire linear regression model, the businesses can enjoy a smaller set of pricing drivers. This becomes particularly handy in the context of communication to the stakeholders, where the number of variables is reduced, and making the pricing choices justified is easier. Nevertheless, stepwise regression is also not always stable when multicollinearity occurs, and it may also fail to detect subtle nonlinear relationships. Although it has a good trade-off between performance and interpretability, more elaborate models such as the Random Forests could be considered when the business is more interested in maximizing its accuracy and feature interaction. The stepwise regression is, hence, a good halfway covenant between straightforward interpretability and predictive stringency.

**Random Forest Regression**

The general objective of the project is to develop a sound and data driven method of forecasting the resale value of used smartphones. Random Forest regression is directly beneficial to this purpose as it uses an ensemble of decision trees to reflect complicated and nonlinear interactions among predictors including RAM, storage, battery capacity, release year, and connectivity features. In comparison with linear models, which presuppose straight-line associations, the Random Forest may represent a complex pattern in the data, being resistant to outliers and multicollinearity. This will not only make the model accurate, but also flexible enough to conform to the different specifications and brand variations that exist in the secondary smartphone market.

Test R 2 and Test RMSE of the Random Forest regression model are 0.860 and 0.226 respectively, the highest of all models tested. This implies that the model accounts for 86% of the variance in the resale prices and it has the least prediction error on unknown data implying good generalization. The ensemble approach is more stable as it averages, across a large number of trees, thus reducing variance and overfitting, relative to a single decision tree. Interpretability is also improved by the features of variable importance measures as they draw attention to the best smartphone features that can predict prices, bridging the gap between accuracy and business insights.

The implication of the Random Forest regression is both important to business and consumer decision-making. The high predictive accuracy of the model makes it the most suitable in cases where the priority would be to reduce pricing errors and guarantee competitive and fair valuations. It is robust enough to manage noisy data and correlated features that are typical of real world smartphone data. The trade-off, however, is less transparency than linear models- the feature importance plots give the direction but the contribution of each attribute itself is less intuitively apparent to explain. Because of this reason, Random Forest is especially usable, when accuracy and competitiveness are of high priority, whereas linear or stepwise regression can still be useful to communicate with stakeholders and have easier interpretability.

**Logistic Regression**

The objective of the classification of logistic regression was to distinguish between high-price (1) and low-price (0) smartphones. This is directly in line with the business objective of assisting the resellers and buyers to make fast and informed choices on whether a device should be placed in the premium resale category or not. Logistic regression is particularly well adapted to this translation of goals in that it generates probabilities of being in a class and therefore permits threshold-based decision-making. In business context, this facilitates the classification process to be interpretable and transparent because the coefficient of each predictor can be traced up to odds of increased resale value.

The logistic regression classifier reached the Accuracy = 87.9% with the Kappa =0.7548 indicating that the predictions that the model produces and the real life performance are highly correlated. The model has a sensitivity (true positive rate) of 0.896 indicating that it has successfully identified close to 90 % of high value smartphones, a fact that is vital in shunning the undervaluation of the premium devices. The specificity of 0.858 signifies that the company could perform better in the segment of identifying low-value devices and this would help prevent overpricing. The balanced accuracy at 0.877 indicates that the model is consistent in both classes with no bias on either of the two categories. All these findings point to the fact that logistic regression has a high predictive and balanced classification power.

A screenshot of a computer

AI-generated content may be incorrect.

This model has two implications. First, logistic regression is business interpretable because it establishes the degree of relationship between annual release year, camera specifications, and RAM with the chances of reselling them, which are easily explained to the stakeholders. Second, the high validity, sensitivity and specificity implies that it can be used operationally with high reliability in the establishment of competitive price categories. There is, however, a risk that complex, nonlinear predictor-predictor correlations can be overlooked by logistic regression as opposed to tree-based ensemble techniques such as Random Forest. Although it is not as efficient in maximizing predictive accuracy, it is still a very good option in case the clarity and justification of decisions are of essence to the extent that performance does not hold as much importance.

**Decision Tree**

A diagram of a decision tree

AI-generated content may be incorrect.

The objective of the decision tree model was to develop a rule based system used to divide smartphones into high and low resale prices. This is directly related to the business requirement of transparency and interpretation of pricing strategies. Decision trees have a strong capacity to apply analytical objectives into practical insights since they can generate clear decision rules e.g. devices with higher normalized new prices, bigger batteries, or more robust camera specifications have higher chances of being categorized as high-value. These kinds of intuitive rules are easy to communicate to resellers and consumers, and they close the waterfall between the outputs of data science and the practical pricing decisions.

A screenshot of a computer

AI-generated content may be incorrect.

The decision tree has a high level of agreement with a predicted versus real category, making the accuracy of the decision tree to be 85.6 % and a Kappa value of 0.707. Sensitivity was 0.893, and it indicated high capability to effectively recognize the high-value devices, which is essential to avoid undervaluing high-quality smartphones. Specificity of 0.810 indicates that the model also does a good job of identifying low-value devices which limits the chances of overpricing. Consistency between the two classes is proven by the balanced accuracy of 85.2% of the classifier. Although a bit less precise than the logistic regression or random forest, the decision tree is competitive, yet it is more interpretable due to its rule-based format.

The important implication of the decision tree is that it is simple to explain and adopt business wise. It gives stakeholders the confidence in the fairness and consistency of pricing decisions because it visually shows how features such as normalized new price, battery size, and rear camera resolution drive classification. Yet, the model is better suited to overfitting as compared to ensemble methods such as Random Forest, and the predictive power may decrease when using more complicated or noisy data. This notwithstanding, decision trees are very useful in situations where clarity is paramount and so they are a good choice in a situation where interpretability is more important than maximum predictive power.

**Random Forest Classifier**

The objective of the classification of the Random Forest model was to correctly classify smartphones into high-price and low-price category and overcome the weaknesses of single decision trees. This matches the business goal of providing very precise but robust classification that is capable of dealing with the variations of smartphone specifications and brand impacts in the resale market. Random Forest does this by averaging the output of the hundreds of decision trees, decreasing variance and increasing generalization. In business terms, this will guarantee that there will be fewer chances of sellers and buyers being misclassified and people will trust in the pricing model.

A screenshot of a computer

AI-generated content may be incorrect.

Random Forest classifier had 87.6 % accuracy and a Kappa of 0.748, which shows that it is in full agreement of the predictions and the real labels. Sensitivity was 0.913, and the performance had excellent powers to identify high-value smartphones correctly, which minimizes the chances of low-valuing premium devices. The model has a specificity of 0.831 that will ensure that it is also effective when it comes to marking devices that have low values, to prevent excessive costs of listings. Balanced performance of 87.2 % is an indication that the model is consistent in the two classes. Such outcomes are better than those of the single decision tree indicating the power of random forest in responding to complex correlations and noisy predictors.

The use of the Random Forest has some major implications to the business in this context. It is highly sensitive, and the resellers are assured that they can recognize the high value devices and gain profitability, and its high specificity ensures that it does not deceive the buyer with expensive models of low value. Interpretability is lower in the trade-off, though, than with logistic regression or a single decision tree. Although feature importance rankings do give a clue about what drives the classification, the inner workings of the model are a bit harder to justify. However, Random Forest offers particularly good results when accuracy and strength are important and is therefore very useful in operational implementation in the volatile used smartphone market.

**Comparison of Regression Models**

Linear regression model was a powerful baseline with Test R 2 of 0.849 and RMSE of 0.235. Its advantage is that it is interpretable; it can give clear coefficients that would be used to determine the effects in which each predictor influences the resale price. This renders it suitable in business applications where transparency is so important. Nevertheless, the model presumes the existence of linear relationships; it is also prone to multicollinearity, which can affect its capacity to explain more complicated relationships among smartphone specifications including battery, RAM, and screen size.

This method was further refined using stepwise regression to only include the most significant predictors, which increased the R 2 slightly to 0.846 with RMSE of 0.238. Stepwise regression simplifies the model, thereby making it easier to understand and concentrate to the most significant features which could allow the stakeholders to make faster and more confident decisions regarding pricing. However, the stepwise selection can miss the subtle or nonlinear patterns and can be non-robust in datasets with correlated predictors and hence is better suited in cases where simplicity and clarity are of importance than marginal gains in accuracy.

Random Forest regression showed the highest predictive accuracy with Test R 2 =0.860 and RMSE= 0.226. It is resistant to outliers and multicollinearity unlike linear-based models and it is efficient in nonlinearities and interactions amongst variables. Interpretability is decreased with this trade-off since it is a black-box ensemble trainer. Nevertheless, the most influential predictors can be determined with the assistance of the feature importance scores. Random Forest is the most powerful tool of all in the business scenario to achieve maximum accuracy and competitive advantage on pricing strategy, yet linear and stepwise regression would still be better suited to business situations where a simple explanation would suffice.

**Comparison of Classification Models**

The logistic regression model was very strong and its accuracy was 87.9, sensitivity (0.896) and specificity (0.858) are balanced. Its key strength is that it can be interpreted as it can explicitly demonstrate the effect of predictors such as RAM, year of release, or connectivity features on the probability of a device to be in a high-price category. This is especially useful where the transparency and explainability are needed, like explaining pricing decisions to the stakeholders. Logistic regression however presupposes linear relationship in the log-odds space thus limiting it to the ability to model nonlinear interaction between variables.

The decision tree model provided an intuitive structure, which is rule-based and achieved the accuracy of 85.6, sensitivity of 0.893 and specificity of 0.810. Although it is a little less accurate than logistic regression, its primary strength is that it can be interpreted. The visual tree illustrates clearly the determination of classification by thresholds in features such as normalized new price, battery capacity and camera resolution. This decision tree is rule based and hence easily adopted in business situations where things need to be explained. Nevertheless, they risk being overfitted, and do not always work better on more complicated data than ensemble techniques.

The most balanced results were obtained with the random forest classifier when the accuracy was 87.6, sensitivity of 0.913, specificity of 0.831, which is better than the single decision tree and the same logistic regression in terms of the overall performance. Ensuring its ensemble design minimizes variation and captures complicated interaction of features, it is more robust and generalizable. The trade-off is less interpreted than logistic regression or decision trees, but feature importance rankings do give some insight. Random Forest is the most dependable algorithm when businesses are more interested in the highest predictive power and soundness of their prices, and logistic regression and decision trees can still be employed where legibility and communication take equal priority.

**Conclusion**

Regression models analysis indicates that although linear and stepwise regression have good interpretability and competitiveness, the regression model with the highest predictive power is the Random Forest regression with R 2 = 0.860 and minimum RMSE =0.226. This shows a significant role played by nonlinear and interaction effects in the price of resale of smartphones and ensemble approaches are in a better position to capture these dynamics. However, linear and stepwise models are still worthy in those cases when business stakeholders would like to be provided with clear explanations on how each of the features like RAM, battery, and release year affects prices.

On the classification side, the results point to a trade off between accuracy and interpretability. The logistic regression and the Random Forest classifier offered high accuracies of around 88 with the Random Forest being slightly more sensitive and robust whilst decision trees offered the most understandable rule based structure albeit with very low accuracy. A combination of these results indicates that businesses with the highest predictive reliability needs should be inclined towards the use of Random Forests, but that those that need clarity and straightforward communication, including reseller justification or customer-facing information, are best served by linear or logistic regression or simple decision trees.

**Recommendations**

To begin with, the Random Forest regression model should be given priority in case good resale prices are to be predicted. Its high performance and highest R 2 as well as the lowest RMSE as compared to the tested models guarantees that business could predict resale values more accurately. This will minimize chances of overpricing low-value devices or underpricing high-value devices and enables resellers to compete well in the high-paced tertiary smartphone market. Random Forest is especially suited to large datasets, whose interactions are complex, so it would work best in operation where predictive accuracy is directly related to profitability.

Second, linear regression and stepwise regression are of great use in the context of business transparency and communications with the stakeholders. The two models have interpretable coefficients to indicate clearly how each of the factors listed in the features (RAM, camera resolution, and release year) affects the price at resale. This is further narrowed down by stepwise regression which only looks at the most effective predictors and hence the non-technical stakeholders can easily interpret and believe the model outputs. The models are specified where clarity rather than maximum accuracy is desired like in reporting or formulation of pricing guidelines.

Third, the most robust is the Random Forest classifier to apply binary classification of devices into price categories. It is also sensitive and specific, which allows maintaining a high accuracy, resulting in reliability when separating high-value and low-value smartphones. This is especially handy in automated systems where machines would require classifying devices more quickly as part of inventory price or online auction sales. Nevertheless, in cases where interpretability is of priority, this can be supplemented with logistic regression or a decision tree, which will give a degree of transparency on how classification decisions are made so that the needs of both technicians and businesses are met.

Lastly, as the hybrid business strategy, it is advised to use Random Forest models in core pricing systems and keep the simple models, such as logistic regression or decision trees, as explanatory tools. This method is both predictive and explainable, which assures good operational performance and at the same time is able to justify the reason behind the decisions to the consumers, auditors or the stakeholders. By combining both ensemble approaches and transparent models, this will enable businesses to have a balance of accuracy, trust, and competitiveness in the resale of used smartphones market.