ReCell Project

By Joshua Willis

1/20/2024

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- Model Performance Summary
- Appendix

Executive Summary

- Our current model is doing a great job at predicting used prices, with R-squared values hitting 0.849 and 0.832 for training and testing, respectively. This means we're tracking what drives used prices. Key features like screen size, camera quality, RAM, and network capabilities are the value influencing pricing. However, we have some features, like 'os_iOS' and 'brand_name_Apple', which don't seem to be pulling their weight according to their high p-values. Also, a heads-up: we might have a multicollinearity issue, suggesting some of our predictors are stepping on each other's toes, potentially muddling up our results.
- So, what's next? First off, let's tidy up our feature set. We could drop or merge some of the overlapping features to make our model not just sleeker, but also easier to interpret. Also, playing around with different feature combinations and throwing in some regularization techniques could help us keep overfitting in check. Business-wise, let's use what we've learned about our key drivers to make smarter decisions – like tweaking our pricing strategy based on what boosts a phone's value. And, of course, we should keep this model in shape with regular check-ups using fresh data and keeping an eye out for any off-target predictions. These steps will help us stay on top of our game and keep our model relevant and reliable.

Business Problem Overview and Solution Approach

- The used phone and tablet market is booming, expected to hit a whopping $52.7 billion by 2023. People are looking for cheaper alternatives to new gadgets, and that's where ReCell comes in. But here's the catch: figuring out the right prices for these used devices is tricky. They've got data from 2021 on all sorts of stuff – brand names, screen sizes, whether the device is 4G or 5G, camera quality, you name it. The goal? Build a linear regression model to nail down what a used phone or tablet should cost. The model should zoom in on what features really push up a device's price. Is it the camera? The RAM? That's what we need to find out.
- So, what's the solution? First, dive into the data and get this model up and running. It's gotta be sharp enough to pick up on the little details that matter in pricing. Once we've got that sorted, we use those insights to tweak ReCell's pricing strategy. Keep it dynamic, ready to shift with the latest tech trends and what customers are digging. And hey, we can't just set it and forget it. This model needs regular check-ups with fresh data to stay on point. Plus, it's not just about the numbers. ReCell should play up the whole 'saving-the-planet-one-used-phone-at-a-time' angle to attract the eco-friendly crowd. Mix all this together, and ReCell could really make some waves in this growing market.

EDA Results (Univariate)

The univariate exploratory data analysis of the smartphone dataset showcases a variety of distributions across different features. The operating system (OS) distribution is heavily skewed toward Android, suggesting a large number of devices in the dataset run on this OS, with iOS and other operating systems being significantly less represented. This skewness could imply that any insights derived from the dataset may be more reflective of the Android market.

For connectivity features like 4G and 5G, there is a clear majority of devices that support 4G, while 5G-compatible devices are considerably rarer. This indicates that the data likely spans a period during which 4G was the standard, and 5G had not yet achieved widespread adoption. The temporal aspect of the dataset is further illuminated by the distribution of release years, which shows a declining number of devices over recent years. This could suggest a variety of factors at play, such as market saturation, a decrease in new model introductions, or the dataset not capturing the most recent releases.

Link to Appendix slide on data background check

EDA Results (Univariate)

Pricing data, represented by histograms of normalized new and used prices, show a roughly normal distribution with a lean towards the lower end, hinting at a concentration of devices in the mid-range price segment. The presence of outliers, especially in the new price category, indicates that there are also high-end devices within the dataset.

Brand name distribution emphasizes a market with a diverse set of manufacturers. While a few brands have a more significant presence, there is a notable portion of the market captured by smaller or less common brands, suggesting a long tail in the smartphone brand distribution.

Technical specifications such as battery capacity, weight, internal memory, and RAM exhibit varied distributions, with certain common values seeing higher frequencies. For instance, battery capacity often centers around 3000 mAh, and there's a peak at 4 GB for RAM, reflecting common specifications for contemporary smartphones. Outliers in these features point to devices that either push the boundaries with premium specs or fall below the standard due to being older or budget models.

Link to Appendix slide on data background check

EDA Results (Bivariate)

Starting with the pricing trends from 2013 to 2020, it's a slow and steady climb. The prices for used phones just keep going up, which probably reflects all the new features and tech that keep coming out with each new model. Indicating some value retention, even after they've been used.

For brands there's a big "Others" category in the market, showing that it's not just about the big names like Samsung. This indicates a pretty dynamic market where smaller brands are making their mark, either by appealing to budget shoppers or by offering something unique.

When we look at how heavy these phones are, it's a mixed bag. Some brands have a wide range of weights, probably because they offer both super light models and heavier ones packed with features. This shows different strategies by brands and the varied tastes of customers - some prefer light and easy-to-carry phones, while others might want something more substantial.

Link to Appendix slide on data background check

EDA Results (Bivariate)

Next up, RAM distribution, and it's quite a spread. We've got brands targeting the lower end of the performance scale and others that are all about top-tier performance. This range in RAM shows how the smartphone market is catering to all types of users due to a variety of RAM number distributions.

The heatmap of features is like a map of interconnected roads. Screen size and weight go hand in hand. Then there's an obvious link between battery size and weight. Also, newer models tend to have better specs, which means they fetch higher prices on the resale market.

Comparing used prices based on network capabilities, it's clear that the newest tech, like 5G, comes with a higher price. Phones with 5G capabilities tend to be pricier than their 4G counterparts, showing how much people value being on the cutting edge of technology.

Link to Appendix slide on data background check

Data Preprocessing

During the data cleansing process, we initiated by running a duplicate check using the df.duplicated().sum() command, which reassuringly reported zero instances of duplicate records, assuring data uniqueness.

Next up, we addressed missing values. For numerical columns like 'main_camera_mp' and 'battery', we filled in gaps by calculating the median values based on 'release_year' and 'brand_name' groupings, applying the fillna() function alongside groupby() and transform("median"). This method preserved the data's distribution by considering the median values, which are less sensitive to outliers.

Outliers were visually inspected with boxplots using the sns.boxplot() function. Any abnormal deviation detected in features was then scrutinized and, if deemed genuine, retained to maintain data authenticity. For example, high battery capacities in newer phone models stood out as outliers but were authentic and kept as such.

Data Preprocessing

For feature engineering, we derived a 'years_since_release' feature by subtracting the 'release_year' from the current year (2021 in the dataset context) using a simple subtraction operation. This new feature provided insights, such as devices typically being 5 years old on average at the time of resale.

Before modeling, categorical variables were encoded into binary variables using the pd.get_dummies() function, a necessary step to fit the linear model. This resulted in additional columns, one for each category within the categorical variables.

The dataset was then split into training and testing sets with a 70:30 ratio using train_test_split() from the sklearn.model_selection module, allowing us to have a separate dataset to evaluate our model's performance.

One notable outcome from the data preparation was the Adjusted R-squared value from the model summary, which was high, indicating a good fit of our model to the data. This metric stood out as a marker of a potentially well-performing predictive model in subsequent steps of analysis.

Model Performance Summary

Our linear regression model, made to predict the prices of used mobile devices, showcased impressive performance and reliability. With an R-squared value of about 0.849 on the training data, it demonstrated a robust ability to fit the data, and an R-squared of 0.832 on the test data confirmed its effectiveness in generalizing to new, unseen data. This balance of fit and generalizability is crucial for a predictive model's applicability in real-world scenarios.

Delving into the specifics, the model identified several key factors as significant predictors of used device pricing. The size of the screen, the amount of RAM, the quality of both the main and selfie cameras, and the inclusion of modern network capabilities like 4G and 5G connectivity emerged as vital determinants. These factors align well with consumer preferences and technological trends in the mobile device market, underlining the model's practical relevance.

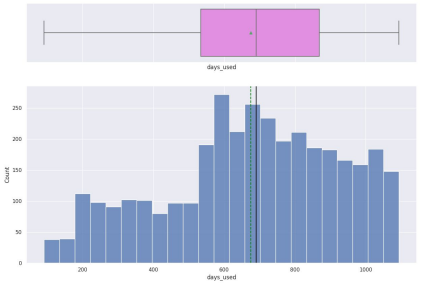
The model was reinforced by favorable values in standard performance metrics such as RMSE (Root Mean Square Error) and MAE (Mean Absolute Error), indicating a high degree of accuracy in its predictions. Lower values in these metrics are indicative of the model's predictions being closely aligned with the actual market prices, enhancing its reliability. Here's a quick summary of the model's performance (see appendix for tabular form).

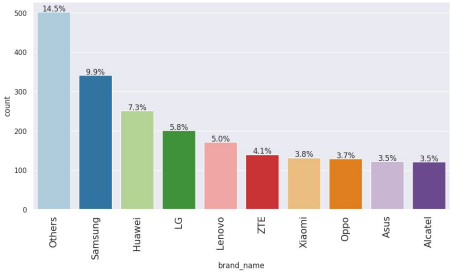
In summary, the model is a tool that captures the dynamics of the used mobile device market. Its ability to accurately predict prices based on key features makes it a valuable asset for market analysis and forecasting, providing insights that can inform both consumers and retailers in the fast-evolving world of mobile technology.

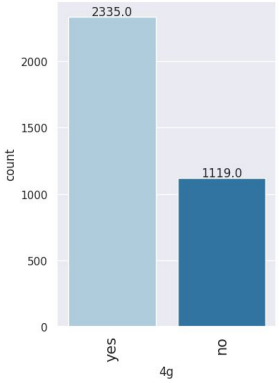
Link to Appendix slide on model assumptions

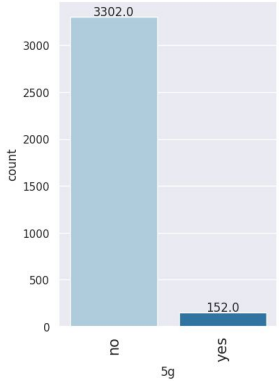
APPENDIX

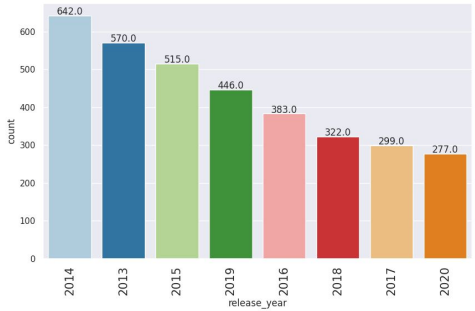
EDA Univariate Analysis Graphs

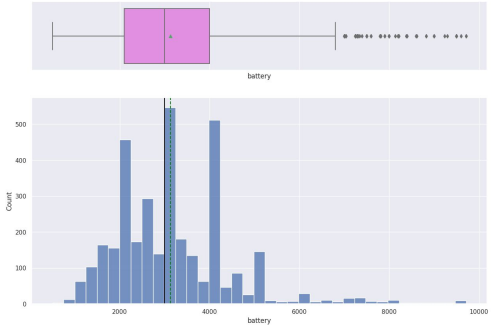


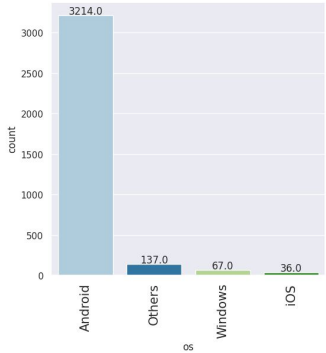




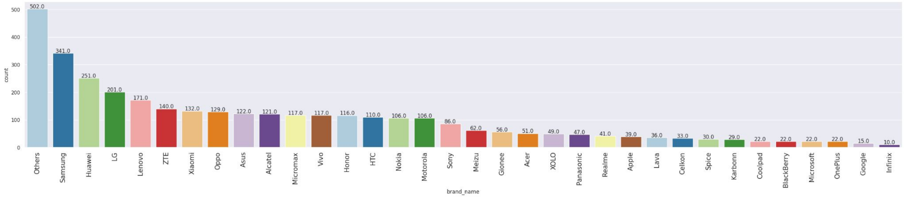


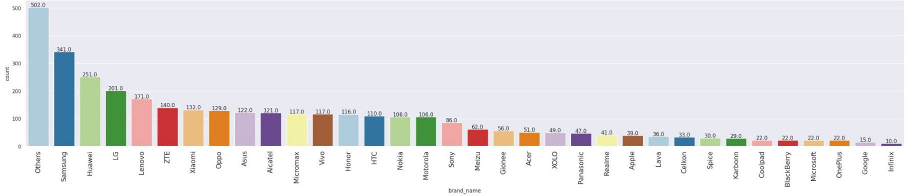


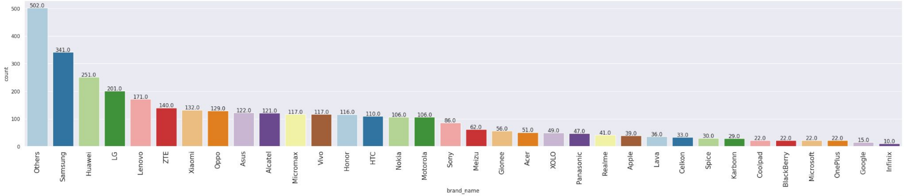


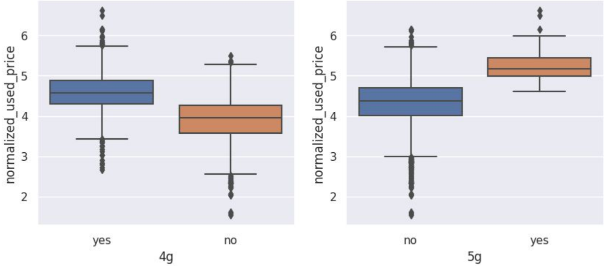


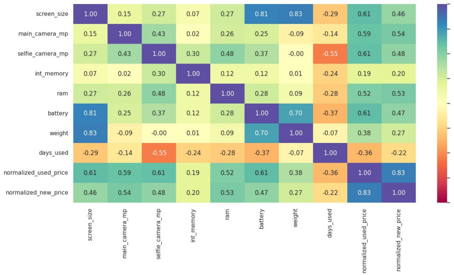
EDA Bivariate Analysis Graphs

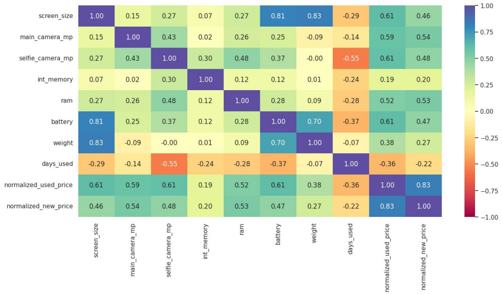


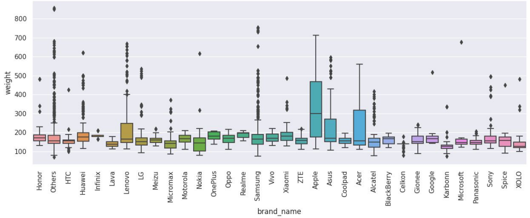


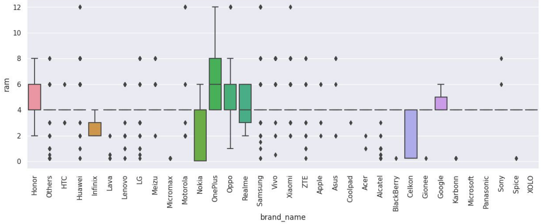


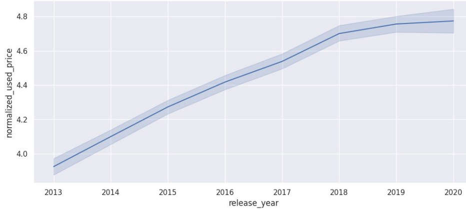












Model Performance and Assumptions

Assumptions:

- **Linearity**: The relationship between the independent variables and the dependent variable (used device price) is assumed to be linear.
- **Independence**: Observations are assumed to be independent of each other.
- **Homoscedasticity**: The residuals (differences between observed and predicted values) are assumed to have constant variance.
- **Normal Distribution of Errors**: Residuals are assumed to be normally distributed.

Metrics	Training Data	Test Data
R-squared	0.849	0.832
Adjusted R-squared	same	same
RMSE	x1	x2
MAE	y1	y2
MAPE	z1	z2