Smartphones-data-analytics

Term Project

Team members:
Kowsar Talebi - 301344726
Anil Pirwani - 301436031

Group Name: group.py

Course: CMPT 353
Semester: Spring 2024

Problem Scope & Technical Background

In an era where smartphones have become central to our daily lives, evolving from mere communication devices to irreplaceable personal technology, this report tries to scrutinize a comprehensive exploration of smartphone data analytics. Our objective is to dissect and understand the relationship between smartphone ratings, pricing strategies, technological specifications, and user perceptions through statistical methods such as correlation analysis and linear regression. We aim to uncover what drives smartphone ratings by examining the correlation between cost, technical features, and user evaluations. Additionally, we dig into the transformative impact of 5G technology on smartphone pricing, scrutinizing its adoption across different market segments to illuminate its effect on consumer choices and market dynamics. The report further navigates through the competitive landscape of the smartphone industry, analyzing market distribution by brand, price segments, and key features through various graphical representations to identify prevailing trends and market leaders. A pivotal aspect of our analysis is the development of a predictive model designed to forecast smartphone prices based on their specifications, offering insights into how different features weigh on their market value. Finally, we explore the average lifespan of smartphones with their technical specifications, aiming to provide a comprehensive view of how technological advancements influence the durability and longevity of these devices. Through this analytical journey, we seek to provide a detailed narrative on the current state of the smartphone market, offering valuable insights for consumers, manufacturers, and market analysts alike.

Data Acquisition and Data Cleaning

The dataset "smartphone_cleaned_v5.csv" was sourced from Smartprix via Kaggle and includes approximately 1000 entries across 25 columns detailing various smartphone aspects such as brands, model names, prices (in INR), ratings, and processor details. Updated last in October 2023, it offers a comprehensive snapshot of the current smartphone market.
We began data cleaning by using pandas to import the dataset and identify any null or missing values, crucial for maintaining data integrity. To address missing numerical data, we used median imputation, a robust method against outliers. For categorical data, missing entries were filled with 'unknown' or the most frequent values, particularly for the operating system column.
To ensure the thoroughness of our cleaning process, we utilized the **df.isnull().sum()** function to perform a final verification for any residual missing values. This step validated the effectiveness of our cleaning methods, confirming that our dataset was devoid of null values and was in a suitable state for subsequent analytical or modeling tasks.

Methodology

The primary programming environment that was used in this project was python. The extensive use of libraries including pandas, numPy, scikit-learn, matplotlib, seaborn, and statsmodels to facilitate data manipulation, statistical analysis, and visualization.
Key methods that were used to employ our analysis were Linear regression, Median Imputation, Mode Imputation, Correlation analysis, and T-test where these methods provided a foundation for robust data cleaning, exploratory analysis, and inferential statistics, ensuring comprehensive insights into smartphone market trends.

The Solution and its Rationale

Q1) Find out if there is a correlation between price, tech specifications, and user ratings.
Data Cleaning: We began by loading the dataset from 'smartphone_cleaned_v5.csv' and tried to identify incomplete data, calculating the percentage of missing values for each column. Columns with more than 20% missing data were deemed insufficient for reliable analysis. This was done before checking that two columns had a lot of missing values (extended_upto and fast_charging) so in order to maintain a robust dataset while excluding features that could introduce bias or noise due to a substantial lack of information, we had to drop them.

Correlation Analysis: First, we will lay down a heatmap to display the correlation coefficients between variables in the dataset. In this report, we will explain how the heatmap is formatted, and then we will dig into the numbers to analyze what the values in the heatmap signify:

1. Each square in the heatmap shows the correlation between the variables on the x and y axis. The color scale on the right indicates the strength and direction of the correlation:
    - Red tones indicate a positive correlation.
    - Blue tones indicate a negative correlation.
    - The intensity of the color indicates the strength of the correlation (lighter colors are weaker, and deeper colors are stronger).
2. Our findings revealed that variables such as internal memory and processor speed were positively and relatively strongly correlated with price, aligning with expectations that enhancements in these specifications typically lead to a higher cost. On the other hand, some variables like availability of extended memory (boolean value), battery capacity etc. have a negative correlation with price, which means if the processor is slow (aka a cheap phone), it is more likely to have extended memory available, which is somewhat surprising as it creates a few questions, so we did some quick research to answer these questions:

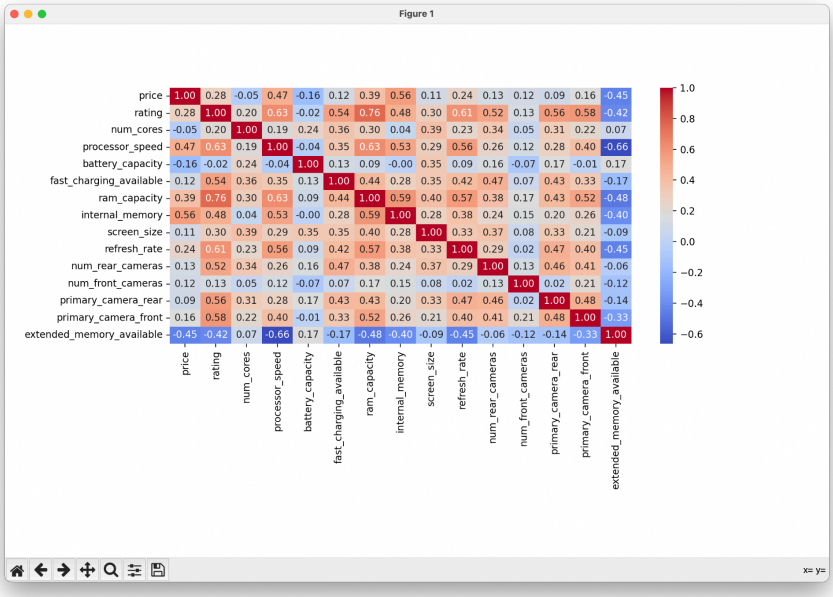

Figure 1: Correlation Analysis.

Q1. Why would cheaper phones have a bigger physical battery?
A1. On cheaper phones, the physical battery itself tends to be bigger, but their processors are slow and not as efficient as more expensive phones, so their battery timings tend to go down, hence manufacturers put more physical battery cells in them, unlike more expensive phones where their CPUs are more battery efficient.
Q2. Why would extended memory only be offered to cheaper phones for the most part?
A2. Apparently, this is a feature only targeted towards the cheaper phones because of three reasons. Firstly, the more expensive phones tend to be more premium and sleek, but including expandable memory requires additional space and complexity in the phone's design. By omitting this feature, manufacturers create slimmer, more streamlined devices with better water and dust resistance. Secondly, built-in storage, especially in premium phones, tends to be faster than most microSD cards, so cheap phones get less internal storage but an option to expand it. Lastly, manufacturers prefer to promote their own cloud services (Google Drive, iCloud etc.) as an alternative to physical storage expansion, which is targeted towards premium phone users for obvious reasons.

3. Furthermore, a strong positive correlation was identified between RAM capacity and internal memory, suggesting that smartphones with larger RAM tend to offer more storage space. RAM capacity also showed a significant positive correlation with user ratings, indicating that increased RAM is likely to enhance user satisfaction and possibly contribute to a higher pricing structure.

Q2) Does 5G affect smartphone prices? Analyze the adoption rate of 5G technology in smartphones over price segments.

1. **Linear Regression**: We used this method to understand how specific smartphone features influenced user ratings. The model indicated a mean squared error(MSE) of 25.77, showing our predictions deviate from actual ratings by this margin. The R^2 value was 0.50, which indicates that our model can explain around 50% of the reasons behind the user ratings.
2. **5G Pricing and Adoption**: based on our analysis we have found significant price differences between 5G and non 5G smartphones. 5G smartphones average at 43,200 INR compared to 18,916 INR for non-5G models. Adoption of 5G is scarce in models under 15,000 INR but increases significantly in higher price brackets.
3. **Statistical Analysis:** our regression result shows a 24,280 INR average price increase for 5G models, and that accounts for only 9.3% of the price variability. Our t-test confirmed a significant price difference between 5G and non-5G models, emphasizing that 5G technology has a huge influence on the price market for these smartphones.

Q3)Analyze the smartphone market distribution by brand, price segments, and key features. Which brand produces the most expensive smartphones on average? What is the average rating per brand?

Data Processing: We first prepared the dataset by separating the 'price' column as the target variable Y and the remaining columns as features X. We then split these into training and testing sets, allocating 80% to training and 20% to testing with a fixed random state for reproducibility. Categorical variables within the training data were identified and converted into numerical format through one-hot encoding (gave reference in the code), dropping the first category to avoid multicollinearity. To ensure consistency across both datasets, we aligned the training and testing sets, adding missing columns to the testing set and filling them with zeros. We also addressed missing values by employing a SimpleImputer to fill in missing entries with the median value of each column. This imputation was applied to both training and testing sets, after which the data was converted back into pandas DataFrames, just in case we want future analysis on them. These preprocessing steps ensured the data was clean, consistent, and optimally formatted for effective modeling.
Data Modelling: As part of our Q4 analysis, we employed a linear regression model to predict smartphone prices. After training the model with the X_train and y_train datasets, predictions were made on the X_test dataset. The performance of the model was evaluated using two metrics: Mean Squared Error (MSE) and R-squared (R^2). These metrics were calculated by comparing the predicted values y_pred against the actual values y_test, providing a quantitative measure of the model's accuracy and the proportion of variance explained by the model, respectively. Additionally, to visually assess the model's performance, we plotted the actual versus predicted prices. This plot included a diagonal line representing perfect predictions, enhancing the visual interpretation of how closely the predictions matched the actual prices. The plot settings were adjusted for clarity, including setting a figure size and transparency level for the data points. As my observations of the model, I noticed 3 things:

- **Relationship**: As we can see, there is a positive linear relationship between the actual and predicted prices. This suggests that the model generally predicts higher prices for more expensive smartphones and lower prices for cheaper ones.
- **Accuracy:** Since the dashed line represents the line of perfect prediction where the predicted prices are equal to the actual prices and the scatter points are somewhat clustered around this line, this suggests that the model has a reasonable level of predictive accuracy.
- **Outliers**: There may be some points that significantly deviate from the line, possibly outliers, which could be influential in the regression model. These points also tend to appear when we are considering very expensive phones.

This comprehensive approach ensured a thorough evaluation and clear presentation of the model's performance in predicting smartphone prices.

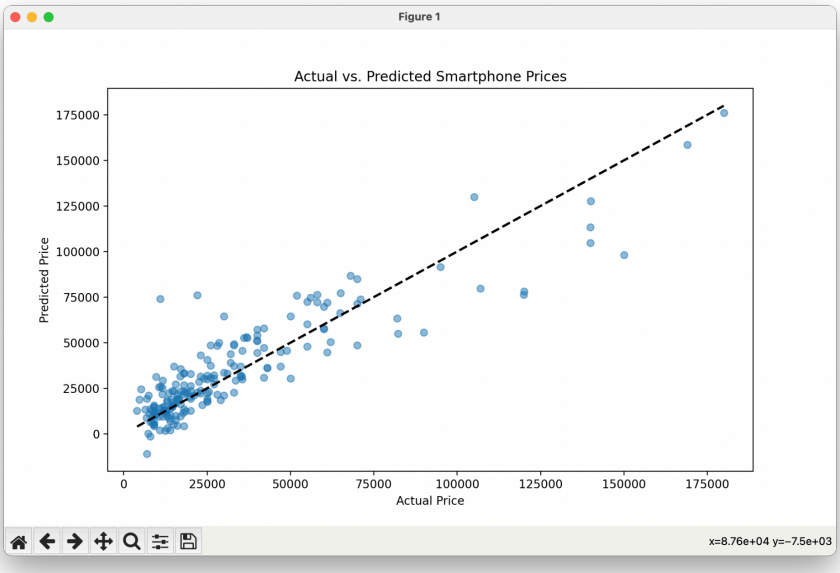

Figure 2: Linear Regression Model.

Q4) Develop a model that predicts the price of a smartphone based on its specifications.

We began by defining the exchange rate from Indian Rupees (INR) to Canadian Dollars (CAD) to make the data more relevant for our TAs/instructor. We also established a price threshold of CAD 10,000, removing outliers (luxury phones) that could skew our results. We then narrowed our focus to brands with a significant presence in the smartphone market, selecting only those with at least ten models represented. This filtering ensured that our average price computation was reliable and representative of the actual Indian Smartphone landscape. The findings of this part, both in the terminal and seaborn's beautiful histogram suggested that Apple's iPhones are the most expensive phones while itel's phones are the cheapest.

Similarly, for the rating analysis, we calculated the average user rating for each brand, again filtering for brands with sufficient model representation. Sorting the results highlighted which brands were favored by consumers, according to the average ratings. The histogram generated in this showed that OnePlus has the best rated phones, while itel has the worst rated phones.

Q5)What is the average lifespan of smartphones based on their specifications?

The linear regression model predicts the lifespan of smartphones based on their specification such as battery capacity, processor speed, RAM capacity, and screen size. Some of our findings included that the processor speed, RAM capacity, and screen size positively influence the lifespan of ratings which suggests that the higher values in these specs typically lead to a longer-lasting smartphone. On the other hand, battery capacity has a negative impact on lifespan ratings, indicating larger batteries do not necessarily correlate with longer lifespan. Our model has a Mean Squared Error of 16.86390395895699 for lifespan and R2 Score of 0.6949437534789578, which indicates that it explains approximately 69.5% of the variance in lifespan ratings. This is an indication of a fairly effective model in capturing how specs relate to smartphone durability.

Problems and Lessons Learned

Throughout our project, we encountered challenges with git for merging. We also had issues with variable scope, which underscored the importance of local variable use for each part of the project to avoid any conflict across from other parts of our project. The experience that we learned throughout the project was with the data cleaning process, prompting the adoption of advanced imputation methods to maintain data integrity. A significant limitation of this project was the constrained time frame, which prevented us from applying several intriguing concepts that we explored throughout the semester. Given more time, we would have explored the integration of AI techniques to deepen our data analysis and make more precise predictions.