**HUMBER COLLEGE OF TECHNOLOGY AND ADVANCED LEARNING**

**Empowering Pricing Strategies**

Introduction

In today's dynamic mobile phone market, firms must accurately forecast the price range of a mobile device in order to compete with companies such as Apple and Samsung. Innovation and product development are required to get through competitiveness. Also, the understanding of user preferences and habits is important for application development. This involves analyzing the relationship between a mobile phone's many properties, such as Random Access Memory (RAM), Internal Memory, and its retail price, which provides insights into efficient pricing strategies. Research says that pricing ranges using machine learning, allowing businesses to make more informed decisions and stand out in an ever-changing market. Through analysis and discovery of important feature-price correlations, we want to help equip organizations to efficiently navigate the market landscape.

Data summary

This dataset gives information about various features of mobile phones, including hardware specifications and characteristics. It includes details such as battery power, Bluetooth support, clock speed, dual SIM capability, front camera resolution, 4G support, internal memory capacity, mobile depth, and mobile weight. Besides this the dataset also contains labels indicating price ranges, which allows us the analysis of the relationship between these features and the selling price of mobile phones.

This dataset is particularly useful for exploring the factors that influence the pricing of mobile phones in the competitive market. By examining the relationship between price ranges and different features we can gain insights into consumer preferences and also make informed decisions about development and pricing strategies.

This dataset contains 21 columns and 1000 rows. We didn’t find any missing values as well as duplicate values in the dataset. Here's a breakdown of each column:

1. ID: Unique identifier for each mobile.

2. Battery Power: Total energy capacity of the battery.

3. Bluetooth: Indicates whether the mobile phone has Bluetooth support.

4. Clock Speed: Speed at which the microprocessor executes instructions.

5. Dual SIM: Indicates whether the mobile phone supports dual SIM cards or not ,

6. Front Camera (FC): Resolution of the front camera.

7. 4G: Indicates whether the mobile phone has 4G network support.

8. Internal Memory: Capacity of internal memory storage.

9. Mobile Depth (m\_dep): Depth or thickness of the mobile phone i.

10. Mobile Weight: Weight of the mobile phone.

11. Number of Cores (n\_cores): Number of cores in the mobile phone's processor.

12. Primary Camera (pc): Resolution of the primary camera.

13. Pixel Height (px\_height): Height of the display screen.

14. Pixel Width (px\_width): Width of the display screen.

15. RAM: Random Access Memory capacity of the mobile phone.

16. Screen Height (sc\_h): Height of the mobile phone's screen.

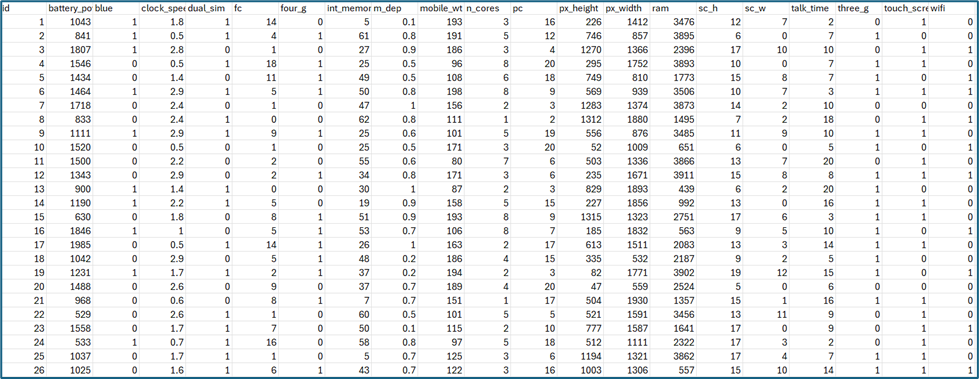
17. Screen Width (sc\_w): Width of the mobile phone's screen.

18. Talk Time (talk\_time): Duration of talk time on the mobile phone's battery.

19. 3G: Indicates whether the mobile phone has 3G network support.

20. Touch Screen: Indicates whether the mobile phone has a touch screen.

21. Wi-Fi: Indicates whether the mobile phone supports Wi-Fi.



**Figure 1: The above table shows the description of the data**

Data Exploration

1. **3G and 4G technology support**

3G network represents the third generation of mobile network technology providing a maximum data bandwidth of 21 Mbps while 4G network Represents the fourth generation of mobile network technology offering a significantly higher maximum data bandwidth of up to 1 Gbps.

A blue and orange pie chart

Description automatically generated

**Figure 2: Pie chart showing 3G network distribution**

The data reveals that most phones in the dataset (76.2%) are compatible with 3G technology, while a minority (23.8%) lack support for 3G.

A blue and orange pie chart

Description automatically generated

**Figure 3: Pie chart showing 4G network distribution**

Slightly more than half of the phones (52.1%) in the dataset support 4G technology while the remaining (47.9%) do not have the same.

After going through this analysis, we can decide whether to consider incorporating 3g or 4g technology in their phones to align with the market demand.

1. **Frequency count for ram**

RAM (Random Access Memory) is like your phone’s short-term memory, which allows it to handle tasks efficiently. With adequate RAM one can manage or handle multiple apps meaning, can switch between the apps without any lag resulting in smooth multitasking and better consumer experience.

A green rectangular box plot with numbers

Description automatically generated

**Figure 4: Boxplot showing the different RAM values**

The dataset shows the RAM values, ranging from approximately **500 MB to 4000 MB. The median RAM value (middle line inside the box) indicates the central tendency.**

1. **Dual sim**

Dual SIM is a feature in mobile phones in which gives an opportunity for the users to use two different SIM cards simultaneously. Each SIM card represents a unique mobile network connection, phone number, and data plan.

A graph of a number of sim cards

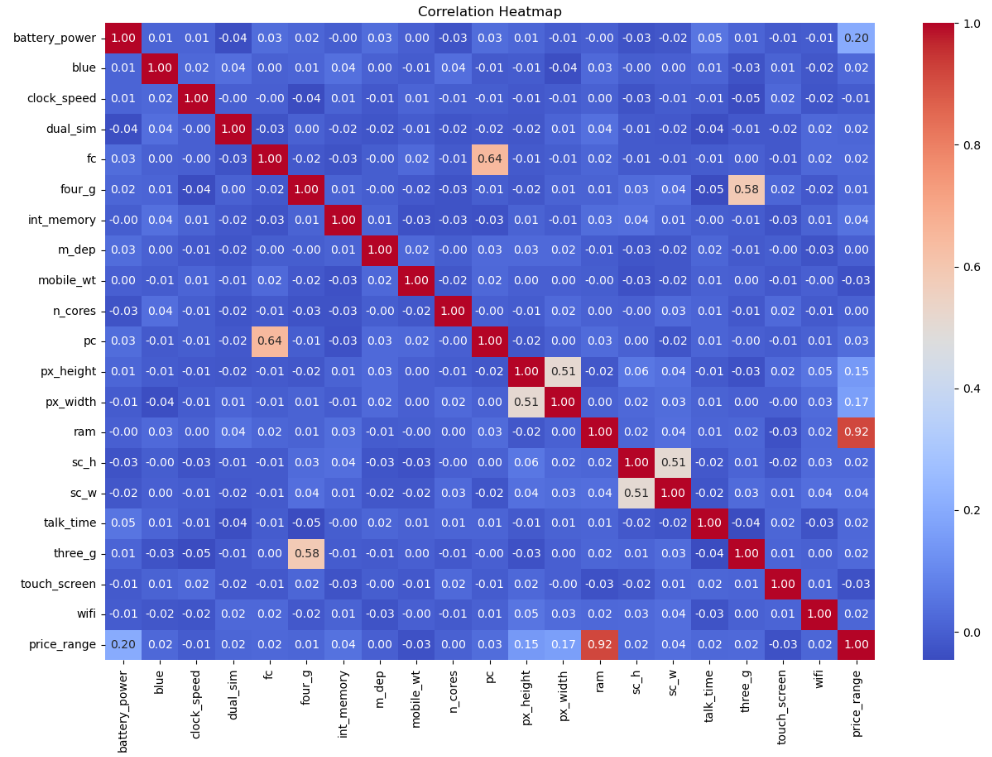
Description automatically generated with medium confidence**Figure 5: Bar Chart showing count of devices that has dual sim feature**

This chart shows the count of devices that support dual SIM functionality versus those who do not. The blue bar represents the count of devices without dual SIM capability (category 0) whereas the orange one represents the count of devices with dual SIM capability (category 1).

This visualization can help understand the market demand for dual SIM devices and guide decisions on whether to include this feature in company or not.

1. **Correlations between variables**

The heatmap helps identify which variables are strongly related and which are not.



**Figure 6: The heatmap showing between variables**

Positive correlations

Battery\_power and ram have a strong positive correlation (close to 1), meaning that as battery power increases, RAM tends to increase, similarly RAM and price range also have a positive correlation of 0.9 which shows as the RAM increases price range also increases. Followed by px\_height and px\_width (pixel height and width), indicating that larger screens tend to have more pixels in both dimensions.

Negative Correlations

‘Clock\_speed’ and ‘price\_range’ have a negative correlation, suggesting that higher clock speeds are associated with lower price ranges. ‘Int\_memory’ (internal memory) and ‘price\_range’ also have a negative correlation.

**Weak Correlations:**

Variables with weak correlations (close to 0) are shown in lighter shades for example, ‘blue’ (Bluetooth support) and ‘price\_range’.

1. **Relationship between RAM and Price Range**

A graph with numbers and lines

Description automatically generated**Figure 7: The scatterplot showing the relationship between RAM and price range**

Each point on the plot corresponds to a mobile phone in the dataset, with its RAM value on the x-axis and its price range on the y-axis. As RAM increases we observe different price ranges .Phones with higher RAM tend to have higher price ranges (positive correlation).

However, there are exceptions, as some phones with low RAM also fall into higher price ranges.The scatter suggests that other factors (besides RAM) influence phone prices.

A diagram of a number of colored boxes

Description automatically generated with medium confidence

**Figure 8: The boxplot showing RAM distribution by Price Range**

The horizontal line represents the median RAM value for that price range. The box itself represents the interquartile range (IQR), which contains the middle 50% of RAM values. As we move from lower to higher price ranges:

The median RAM increases. The IQR widens, suggesting greater variability in RAM within each price range. Some outliers may exist, especially in higher price ranges.

1. **Relationship between Mean Battery Power by Price Range**

A graph of battery power

Description automatically generated**Figure 9: The bar graph showing the Mean Battery Power by Price Range**

The bar graph shows the relationship between the mean battery power and different price ranges. The x-axis represents the “Price Range” with three categories: 1, 2, and 3. The y-axis represents the “Mean Battery Power” with values ranging from 0 to 1400.

Each bar in the graph corresponds to a price range. The height of the bar represents the mean battery power for that price range. From the graph, it’s clear that as the price range increases, the mean battery power also increases. This suggests that phones in a higher price range tend to have a higher mean battery power.A higher battery capacity or power often means a more expensive phone. This is because higher battery power leads to longer battery life, which is a desirable feature for many users. Therefore, manufacturers might charge more for phones with higher battery power.

1. **Feature importance for Predicting Price Range**

A graph with blue bars

Description automatically generated **Figure 10: The horizontal bar graph showing the Mean Battery Power by feature importance for Predicting Price Range**

This graph depicts the feature importance for predicting the price range. This kind of visualization is often used in machine learning to understand which features are most influential in a model’s predictions. The graph has multiple bars, which represent different variables in the given dataset. The length of each bar will indicate the importance of that variable in the price range.

From the graph depicted, we can observe that "RAM" is the most significant in the prediction of price range, as the length of its bar is the longest. Other features like ‘battery\_power’, ‘px\_width’, ‘px\_height’, 'int\_memory', etc. also contribute to the prediction, but not as significantly as ‘RAM’. We also observe that the least important factor that affects the price range is "three\_g."

PREDICTIONS

A close-up of a graph

Description automatically generated

**Figure 11: The histogram shows the actual prices Figure 12: The histogram shows the predicted price**

The left histogram represents the actual prices of mobile phones.

A peak around the lowest price range (close to 0), indicating a high frequency of low-priced phones. A smaller peak around the second price range (around 1), suggesting moderate frequency in that range. Another peak around the highest price range (around 3), indicating a significant number of higher-priced phones. The distribution is positively skewed, with more phones in lower price ranges.

The right histogram represents the predicted prices of mobile phones. The distribution is more uniform across different price ranges. There is still a significant frequency in the lowest price range (around 0.5). As predicted prices increase, the frequency gradually decreases. The model seems to predict a wider range of prices compared to the actual distribution.

A graph with blue lines and text

Description automatically generated with medium confidence

**Figure 13: The histogram shows the distribution of predicted price ranges**

The histogram visually represents the distribution of predicted price ranges for mobile phones. It helps us understand how frequently different predicted price ranges occur in the dataset. A peak around the lowest predicted price range (around 0.5), indicates a high frequency of phones predicted to be low-priced. As predicted prices increase, the frequency gradually decreases. The model predicts a wider range of prices, including mid-range and higher-priced phones.

Test training data description

We split our data into two parts to build and test machine learning models. We split the data into training and testing sets making sure the split is fair and consistent each time we do it. Our results show how many samples are in each set which helps us see how model will perform. It is important to split the data this way to make sure our model does not learn too much from training data and can perform well on new unseen data.

We extracted the feature variables from data frame. It removes the column labeled ‘price\_range’ along with the column axis leaving only the feature variables in x. Then we extracted the target variable from data frame which selects only the column label ‘price\_range’ assigns it to variable y. We split the feature and target variables into training and testing dataset through train\_test\_split function which randomly splits the data. We used 33% of the data for testing and 67% of the data for traing. We set the random seed for reproducibility. Setting a random seed ensures that the same random splits are generated each time the code is run, making the results reproducible. After executing the code, we got four datasets:

X\_train: Feature variables for the training set.

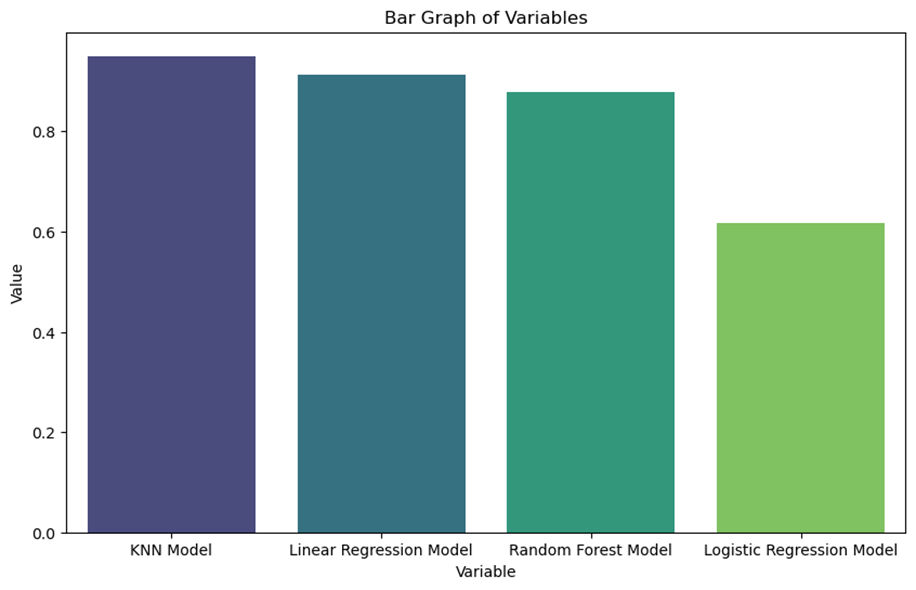
X\_test: Feature variables for the testing set.

y\_train: Target variable for the training set.

y\_test: Target variable for the testing set.

Model performance:

Model performance comparison:



**Figure 14: The bar graph shows the comparison among the models**

The purple bar shows the K-Nearest Neighbors (KNN) model. It has the greatest value and is closest to 1. KNN may be applied to regression and classification problems. It first locates the K data points in the training set that are closest to the input data point. Next, predictions are made using values of those nearby points. When the KNN model's value is near to 1, it indicates that it is functioning well on the assessment measure, which probably means it has high accuracy or predictive capacity.

The Linear Regression Model, shown by the dark blue bar, has lower value than the KNN model. Linear regression models the connection between a dependent variable and one or more independent variables. When the lower score is compared to KNN indicates that the linear regression model performs slightly less accurately, but it is a effective model.

The teal bar shows the Random Forest Mode. It has a value similar to but slightly lower than Linear Regression. The similar result to linear regression indicates that the Random Forest model's performance is same, but lower, which might imply that it is less successful on this specific dataset.

Logistic Regression Model shown by the light green bar, has a value of around 0.6. The value of roughly 0.6 suggests reasonable performance when compared to the other models discussed. While not as good as KNN, it still shows that the Logistic Regression model can produce relatively accurate predictions.

Based on this, the K-Nearest Neighbors (KNN) model happens to be the best choice among the models evaluated. It consistently outperforms the others, achieving the highest value close to 1. This suggests that it's the most accurate model for making predictions based on the given evaluation criteria. Therefore, if accuracy is the primary concern, the KNN model would be the preferred option.