MOBILE PRICE RANGE CLASSIFIER

A Machine Learning Approach to Informed Consumer Choices.

PROJECT BY : JOY ACHIENG OGUTU

Deployed app link here

INTRODUCTION

In a world where technology constantly evolves, choosing the right mobile phone that meets both your needs and budget can be a daunting task. With countless options available, each boasting a myriad of features and price points, consumers often find themselves overwhelmed and uncertain about which device to invest in.

This project aims to alleviate this dilemma by harnessing the power of machine learning to develop a predictive model capable of classifying mobile phones into different price ranges based on their unique attributes. By doing so, we strive to provide consumers with a valuable tool that empowers them to make informed purchasing decisions.

PROJECT OVERVIEW

Business Understanding

The rapid evolution of technology has led to an explosion of mobile phone models with varying specifications and price points flooding the market. With such a wide array of choices, consumers often find themselves grappling with the dilemma of selecting a device that not only meets their requirements but also fits within their budget. This challenge is exacerbated by the lack of clear guidance and information, leaving consumers vulnerable to making uninformed decisions or being swayed by marketing tactics.

Problem Statement

We aim to address the pressing issue of consumer uncertainty and confusion when navigating the mobile phone market. Through data-driven insights and predictive analysis, our goal is to empower consumers with the knowledge they need to make confident and informed purchasing decisions.

Objectives

The primary objective of this project is to develop a machine learning model capable of accurately classifying mobile phones into predefined price ranges based on a diverse set of attributes. Other objectives include:

- To explore and preprocess the dataset to handle missing values, outliers, and any other data inconsistencies.
- To perform exploratory data analysis (EDA) to gain insights into the relationships between different features and the target variable (*price_range*).
- To select appropriate machine learning algorithms for classification and evaluate their performance using suitable metrics.
- To fine-tune the chosen model to improve its predictive accuracy.

- To validate the final model using cross-validation techniques to ensure its robustness.
- To deploy the model for predictions.

DATA UNDERSTANDING

The dataset provided contains a comprehensive set of features describing various aspects of mobile phones, including battery power, camera specifications, memory capacity, connectivity options, and more. Each mobile phone entry is labelled with its corresponding price range, ranging from low to very high cost.

The dataset comprises of the following columns:

- *battery_power* - Total energy a battery can store in mAh.
- *blue* - Bluetooth enabled (1 if yes, 0 if no).
- *clock_speed* - Speed at which a microprocessor executes instructions.
- *dual_sim* - Dual SIM support (1 if yes, 0 if no).
- *fc* - Front Camera megapixels.
- *four_g* - 4G network support (1 if yes, 0 if no).
- *int_memory* - Internal Memory (in gigabytes).
- *m_dep* - Mobile Depth in cm.
- *mobile_wt* - Weight of mobile phone.
- *n_cores* - Number of cores of the processor.
- *pc* - Primary Camera megapixels.
- *px_height* - Pixel Resolution Height.
- *px_width* - Pixel Resolution Width.
- *ram* - Random Access Memory in megabytes.
- *sc_h* - Screen Height of mobile in cm.
- *sc_w* - Screen Width of mobile in cm.
- *talk_time* - Longest time that a single battery charge will last when you are talking.
- *three_g* - 3G network support (1 if yes, 0 if no).
- *touch_screen* - Touch screen support (1 if yes, 0 if no).
- *wifi* - Wifi connectivity (1 if yes, 0 if no).
- *price_range* - Price range of the mobile phone (0 - low cost, 1 - medium cost, 2 - high cost, 3 - very high cost).

This rich dataset serves as the foundation for our machine learning model, enabling us to explore the relationships between different features and price ranges, ultimately leading to the development of an accurate classification system.

DATA PREPARATION

Data preparation is a crucial stage in this project for a number of reasons:

- *Feature Engineering*: New features might be developed or current ones modified in order to improve analysis.
- *Handling Missing and Duplicate Data*: Analysis results can be greatly impacted by missing and duplicate data. In order to achieve a robust analysis, handling missing values must be decided, whether through imputation, deletion, or other suitable approaches.
- *Outlier detection*: For statistical validity, outliers must be found and dealt with. We can use methods like visual inspection or statistical testing to find outliers and handle them correctly with the help of data preparation.

In the data preparation phase, several important actions were taken to ensure the dataset was ready for exploratory data analysis (EDA) and subsequent modelling:

1. **Duplicate Values**
   There are no duplicates in the dataset.

2. **Null Values**
   All columns have no null values.

3. **Outliers**
   Only three columns had outliers, *fc* with 18, *px_height* with 2 and *three_g* with 477. The outliers were retained. Retaining outliers guards against bias being introduced into the analysis, as extreme values may represent genuine data points rather than errors.

   Additionally, the presence of outliers contributes to the robustness of analysis, as it ensures that statistical methods and machine learning algorithms are not unduly influenced by extreme values.

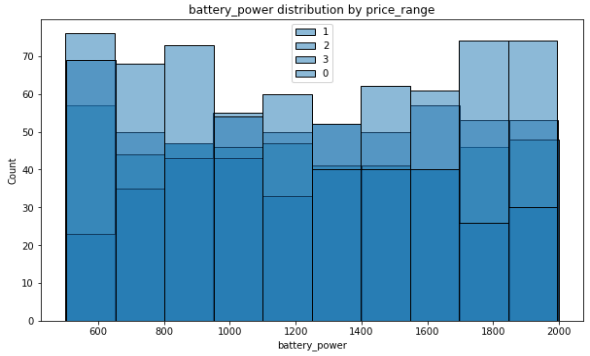
   Lastly, analysing outliers aids in identifying data quality issues such as measurement errors or data entry mistakes, facilitating improvements in data collection processes and overall data quality. Thus, while the decision to keep or remove outliers depends on the specific context and objectives of the analysis, retaining outliers is often essential for ensuring accurate and comprehensive data analysis.

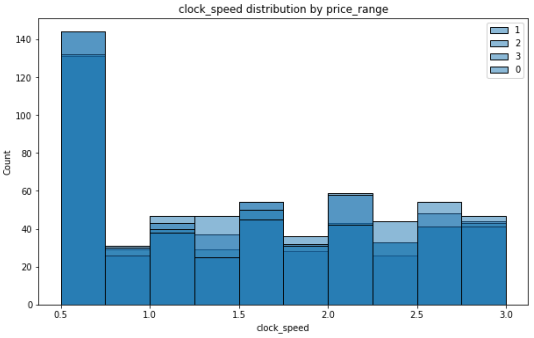
4. **Data Type Conversion**
   The following categorical columns (*blue*, *dual_sim*, *four_g*, *three_g*, *touch_screen*, *wifi* and *price_range*) were converted to category type.

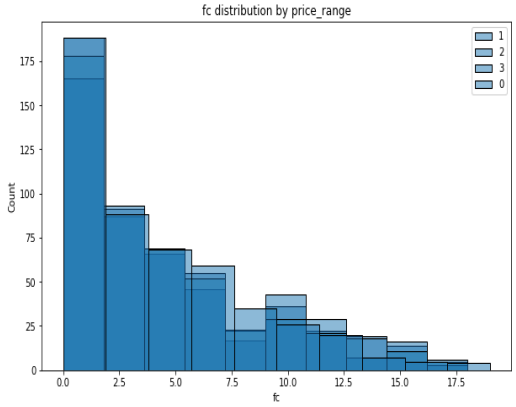
EXPLORATORY DATA ANALYSIS

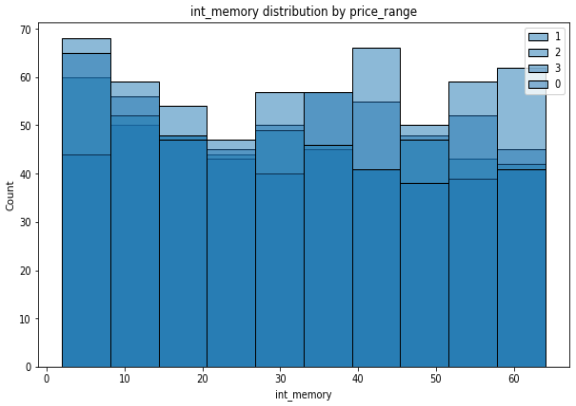
A crucial turning point in our investigation occurs during the data analysis stage, which enables us to extract significant patterns and insights from the compiled dataset. We used univariate, bivariate, and multivariate exploratory data analysis (EDA) methodologies in a comprehensive manner to properly analyse the data.
Let's dive into our analysis.

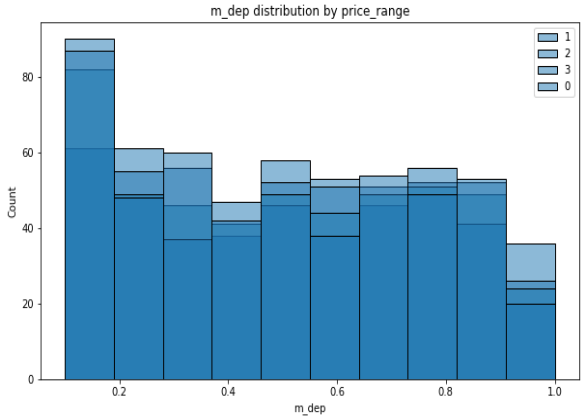
Distribution of both the Numerical Columns and the Categorical Columns

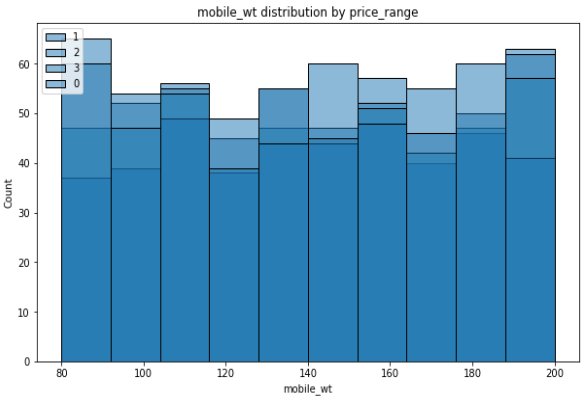


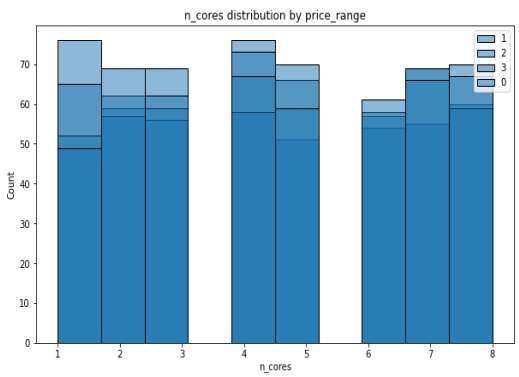


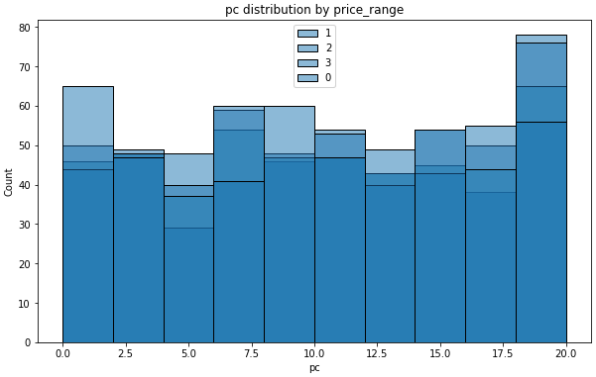


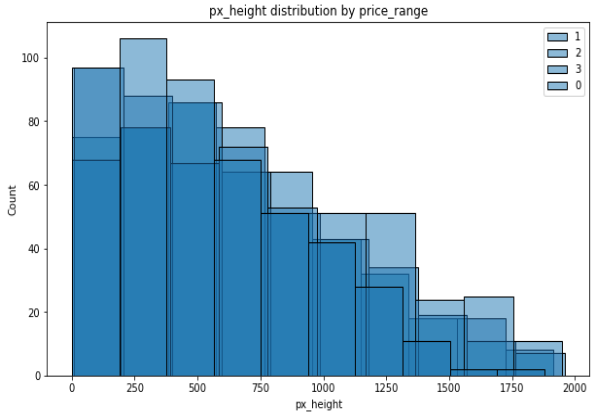


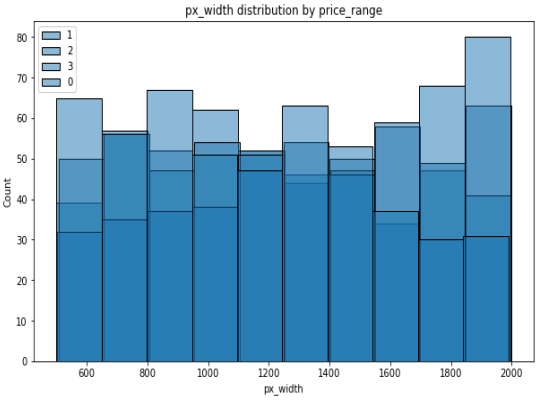


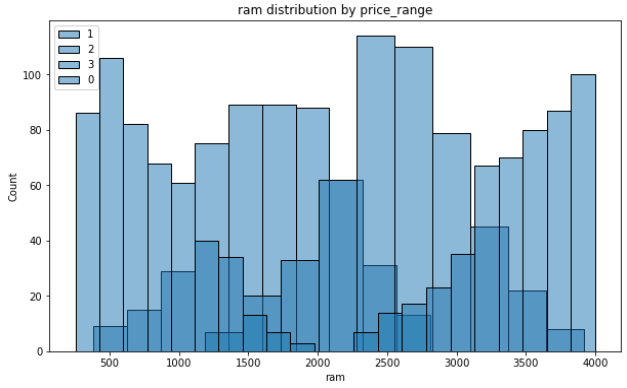


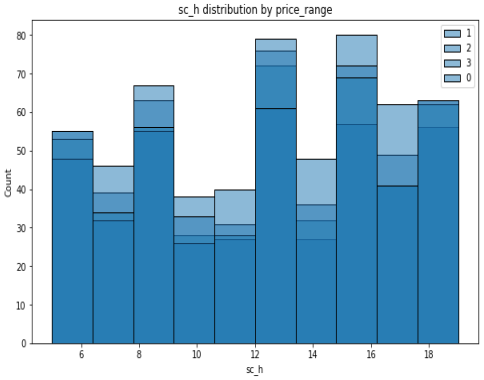

n_cores distribution by price_range
1
2
3
0
Count
n_cores

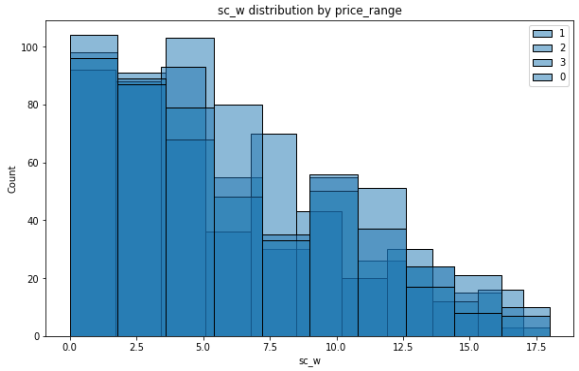

pc distribution by price_range
1
2
3
0
Count
pc

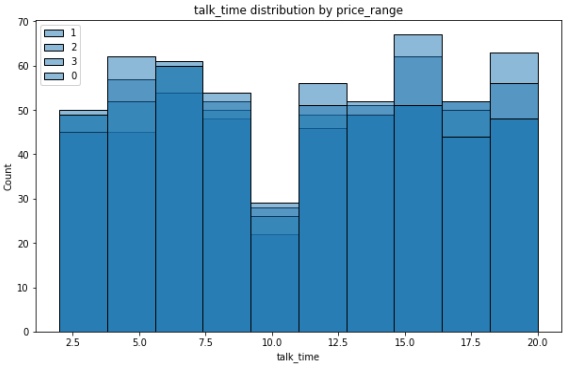

px_height distribution by price_range
1
2
3
0
Count
px_height


px_width distribution by price_range
1
2
3
0
Count
px_width


ram distribution by price_range
1
2
3
0
Count
ram


sc_h distribution by price_range
1
2
3
0
Count
sc_h


sc_w distribution by price_range
1
2
3
0
Count
sc_w


talk_time distribution by price_range
1
2
3
0
Count
talk_time

Histogram Summary:
Histograms provide a visual summary of the distribution of data, allowing users to quickly understand the central tendency, spread, and shape of the data. They are useful for identifying patterns, outliers, and potential anomalies within a dataset. Histograms are particularly effective for exploring the frequency distribution of continuous variables.
Key Components of a Histogram:

- Bins: The intervals or ranges into which the data is divided. Each bin represents a specific range of values.
- Frequency: The count of data points falling within each bin.
- X-axis: Represents the range of values or intervals of the data.
- Y-axis: Represents the frequency or count of data points within each bin.

The histograms for the mentioned numerical columns mostly reveal a right-skewed distribution. In a right-skewed distribution, also known as positively skewed, the majority of data points are clustered on the left side, and the tail extends towards the right.

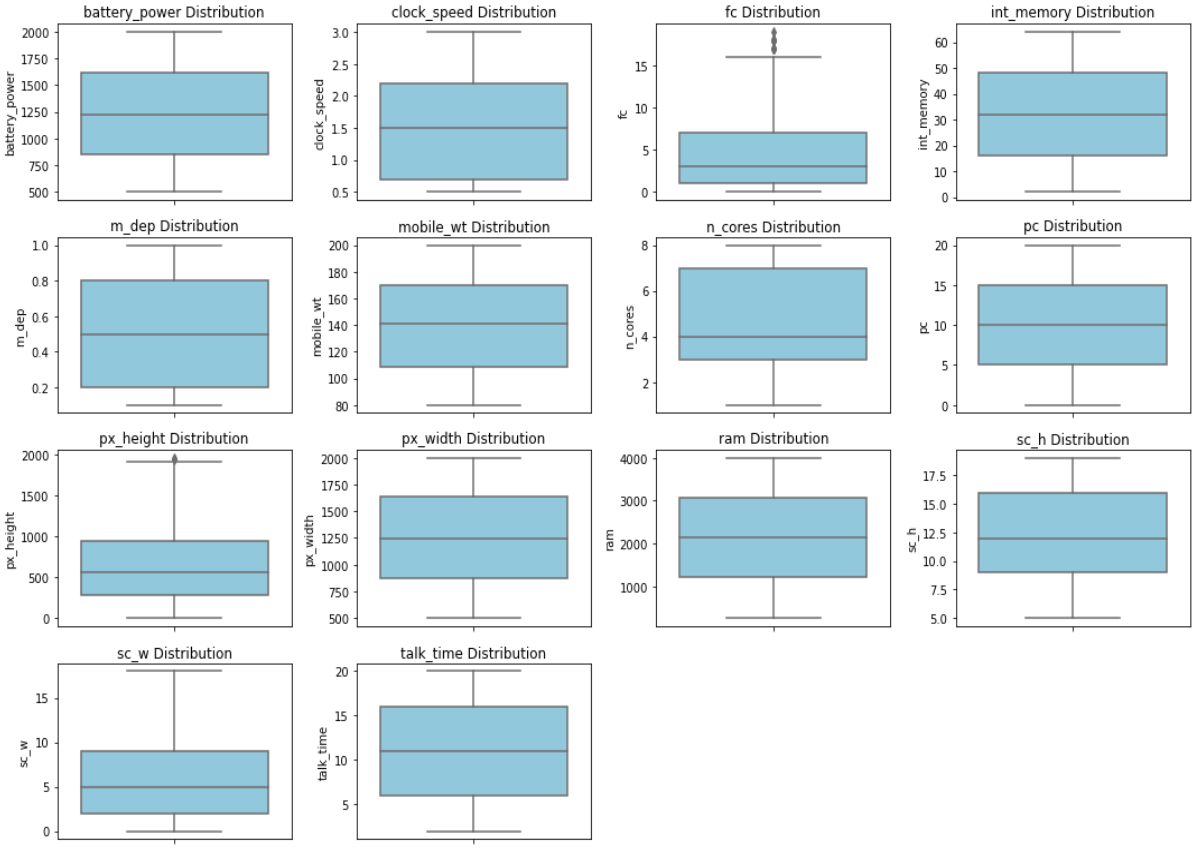
Countplot Summary:
A countplot is a type of bar plot that displays the counts of observations in each category of a categorical variable. It represents the frequency of occurrences of each category within the dataset.

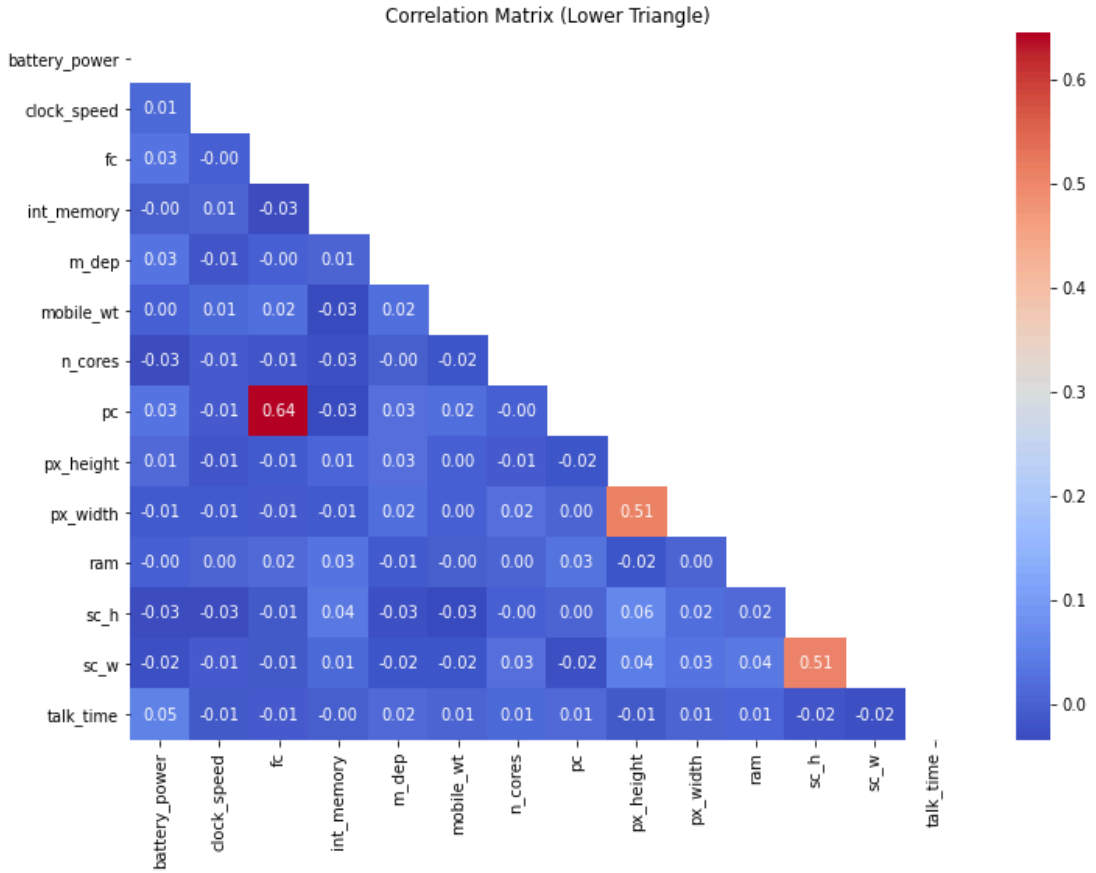
Here's a summary of the above plot:

1. **Representation**
    - Vertical or horizontal bars represent the count or frequency of each category.
    - The height (or length) of each bar corresponds to the number of occurrences of the category it represents.
2. **Interpretation**
    - Higher bars indicate more frequent occurrences of the corresponding category
    - Lower bars indicate less frequent occurrences.
    - Patterns in the distribution of categories can provide insights into the underlying data.

From the count plots above, we can conclude that mobile phones that have the categorical features enabled or present - fall under `high cost` or `very high cost`.

In summary, count plots are a simple yet powerful tool for visualising the distribution and frequency of categorical variables in a dataset, making them valuable for gaining insights into the data's characteristics and patterns.


battery_power Distribution
clock_speed Distribution
fc Distribution
int_memory Distribution
m_dep Distribution
mobile_wt Distribution
n_cores Distribution
pc Distribution
px_height Distribution
px_width Distribution
ram Distribution
sc_h Distribution
sc_w Distribution
talk_time Distribution


Correlation Matrix (Lower Triangle)
battery_power
clock_speed
fc
int_memory
m_dep
mobile_wt
n_cores
pc
px_height
px_width
ram
sc_h
sc_w
talk_time

Boxplot Summary:

The above plot shows the following key statistics of the numerical columns in the data:

1. **Median (Q2)**: The middle value of the columns when arranged in ascending order. It represents the 50th percentile of the data.
2. **Quartiles (Q1 and Q3)**: The dataset is divided into four equal parts, with each part representing 25% of the data. Q1 represents the 25th percentile (lower quartile), and Q3 represents the 75th percentile (upper quartile).
3. **Interquartile Range (IQR)**: The range between the first (Q1) and third (Q3) quartiles. It covers the middle 50% of the data.
4. **Whiskers**: Lines extending from the box that represent the minimum and maximum values within 1.5 times the IQR from the first and third quartiles, respectively. Any data points outside this range are considered outliers and are plotted individually.
5. **Outliers**: Data points that fall outside the whiskers, indicating they are unusually high or low relative to the rest of the data.

Boxplots are useful for identifying the central tendency, spread, and skewness of a dataset, as well as detecting the presence of outliers. They provide a visual summary of the distribution without assuming a specific underlying probability distribution.

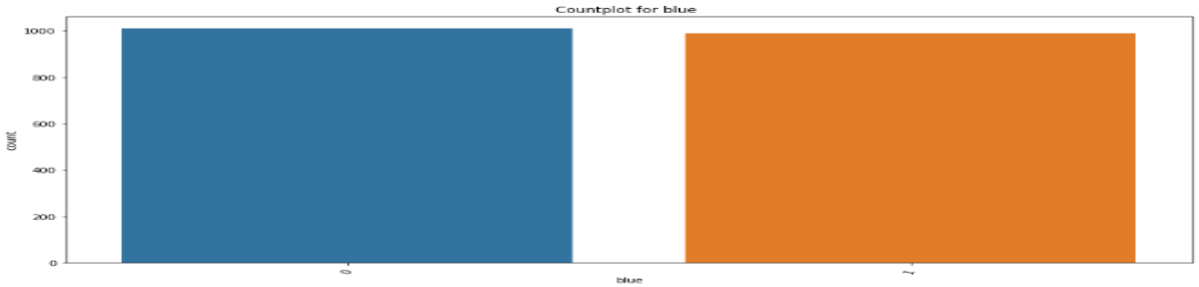
Correlation Heatmap Summary:

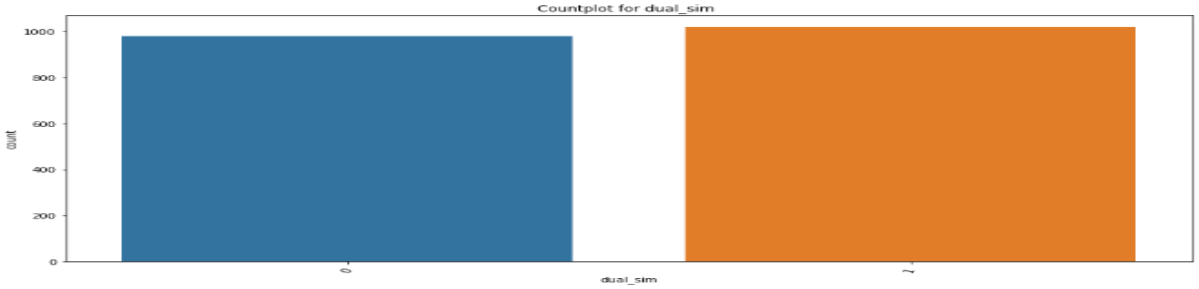
A square matrix where each cell represents the correlation coefficient between two variables. Typically, the correlation coefficient ranges from -1 to 1, indicating the strength and direction of the relationship between variables. Values close to 1 indicate a strong positive correlation, values close to -1 indicate a strong negative correlation, and values close to 0 indicate little to no correlation.

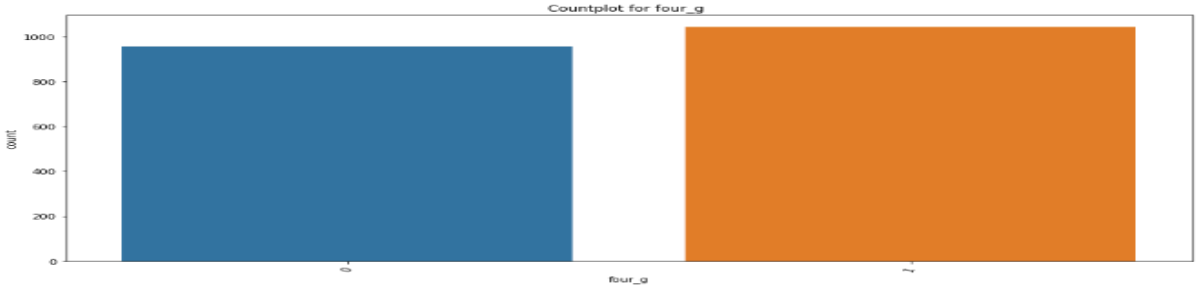
Here's a summary of the above heatmap matrix:

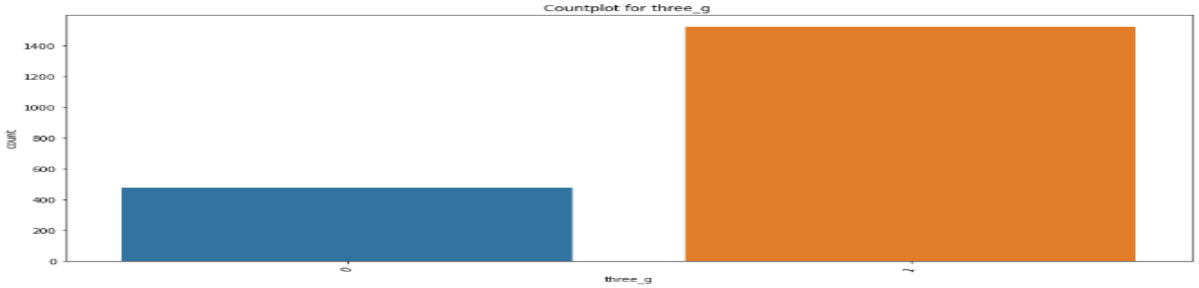
1. **Interpretation**:
   Positive correlation: Variables with a correlation coefficient close to 1 move in the same direction. For example, as one variable increases, the other tends to increase as well. Negative correlation: Variables with a correlation coefficient close to -1 move in opposite directions. For example, as one variable increases, the other tends to decrease. No correlation: Variables with a correlation coefficient close to 0 show no linear relationship.
2. **Visualisation**:
   The correlation matrix above has been visualised as a heatmap, with colours representing the strength and direction of correlations. Warm colours (e.g., red) indicate positive correlations, while cool colours (e.g., blue) indicate negative correlations.

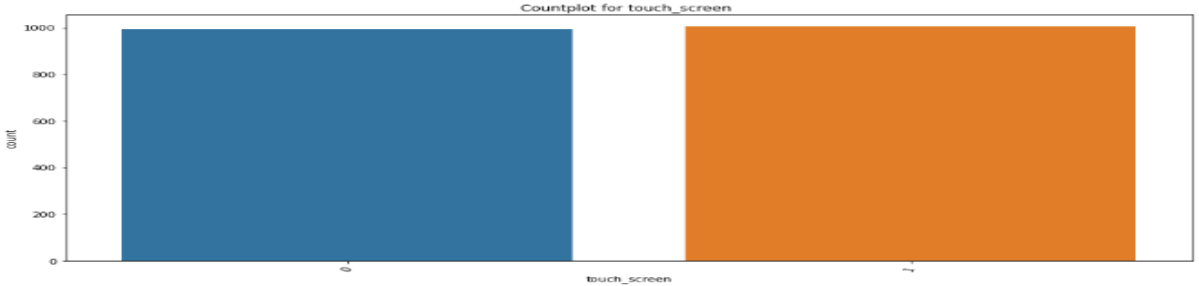
In summary, `pc` and `fc` have a close to high correlation coefficient of 0.64 and `sc_h` and `sc_w` have a correlation coefficient of 0.51.

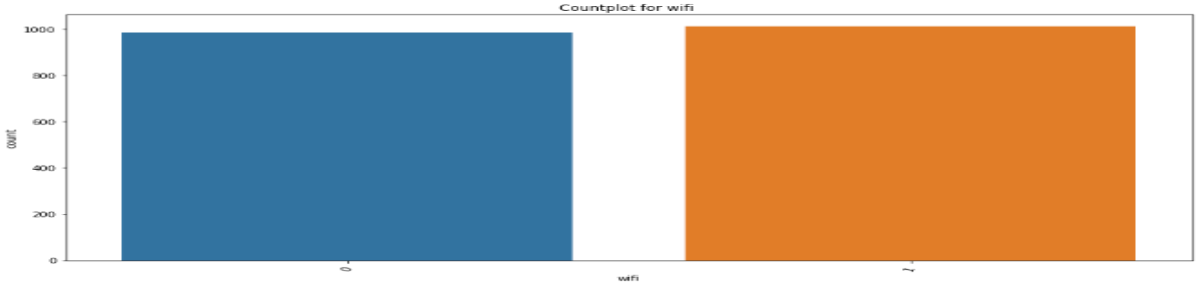

Countplot for blue
count
blue

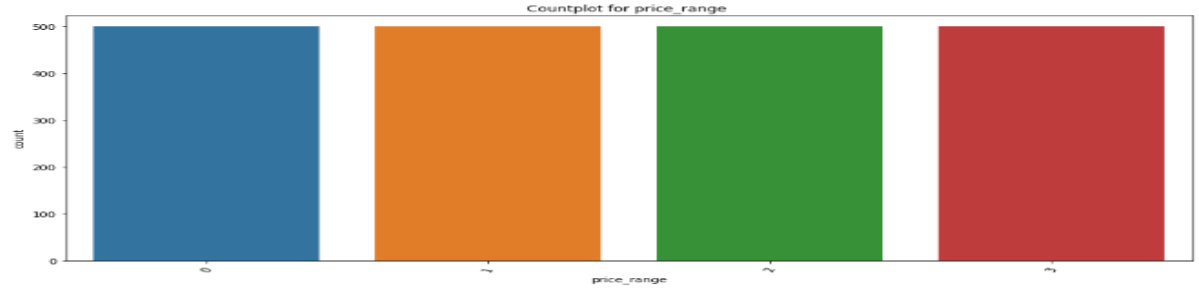

Countplot for dual_sim
count
dual_sim


Countplot for four_g
count
four_g


Countplot for three_g
count
three_g


Countplot for touch_screen
count
touch_screen


Countplot for wifi
count
wifi


Countplot for price_range
count
price_range

MODELLING

Modelling a multiclass dataset involves building predictive models to classify observations into multiple categories or classes. Unlike binary classification, where there are only two possible outcomes, multiclass classification deals with scenarios where there are three or more distinct classes to predict.

In this project, we delve into the complexities of modelling a multiclass dataset, aiming to develop robust algorithms capable of accurately assigning observations to their respective classes. By leveraging machine learning techniques, we seek to uncover patterns and relationships within the data that can aid in effective classification.

We will evaluate the model with accuracy. The accuracy metric is well-suited for evaluating this model for several reasons:

1. **Interpretability:** Accuracy provides a straightforward measure of how often the model correctly predicts the mobile phone price range. It represents the proportion of correct predictions out of the total predictions made by the model.
2. **Intuitive Interpretation:** In the context of mobile price classification, accuracy directly translates to the percentage of mobile phones that are correctly classified into their respective price ranges. This intuitive interpretation makes it easy for stakeholders and end-users to understand the model's performance.
3. **Balanced Performance:** Accuracy considers both true positives and true negatives, making it suitable for balanced datasets where the classes are evenly distributed. It provides an overall assessment of the model's effectiveness across all classes, which is important in multiclass classification tasks like this one.
4. **Simplicity:** Accuracy is a simple and easy-to-understand metric, which makes it accessible to stakeholders with varying levels of technical expertise. Its simplicity facilitates effective communication of model performance and results.

While accuracy is a valuable metric for evaluating model performance, it's essential to consider other metrics as well, especially in scenarios where class imbalance or misclassification costs are significant. Depending on the specific requirements and objectives of the project, metrics such as precision, recall, F1-score, and confusion matrix analysis may provide additional insights into the model's performance.

One-vs-Rest (OvR) Multiclass Strategy

When dealing with a multiclass classification problem, the one-vs-rest (OvR) strategy allows logistic regression to train a separate model for each class, comparing it against all the other classes. Also known as the One-vs-All (OvA) strategy, this method creates one logistic regression model for each class against the rest of the classes in the dataset.

For example, our target vector contains four classes (0, 1, 2, 3), the OvR strategy will create four separate models as follows:

- Model a: 0 vs [1, 2, 3]
- Model b: 1 vs [0, 2, 3]
- Model c: 2 vs [0, 1, 3]
- Model d: 3 vs [0, 1, 2]

One-vs-One (OvO) Multiclass Strategy

In contrast, the one-vs-one (OvO) strategy trains a separate model for each class against every other individual class in the dataset. This results in (n * (n-1) / 2) models, where (n) represents the number of classes.

For instance, in a scenario with classes 0, 1, 2, and 3, the OvO strategy will create six separate models:

- Model a: 0 vs 1
- Model b: 0 vs 2
- Model c: 0 vs 3
- Model d: 1 vs 2
- Model e: 1 vs 3
- Model f: 2 vs 3

Multinomial Method

In the multinomial method, logistic regression employs the softmax function instead of the logistic function used in OvR and OvO strategies. This function calculates the probability value of each class over (n) different classes.

Choosing the Best Strategy:

- OvR is computationally efficient and provides easily interpretable models.
- OvO may yield higher accuracy but is computationally expensive due to its (O(n^2)) complexity.

Multinomial method offers reliable probability estimates but requires more complex optimization algorithms.

For this project, considering the classes [0 - low cost, 1 - medium cost, 2 - high cost, 3 - very high cost], the default OvR strategy may be suitable due to its simplicity and interpretability. However, experimenting with different strategies is recommended to determine the most effective approach for the specific dataset.

MODEL	Accuracy
One vs Rest (OvR)	79.5
One vs One (OvO)	73.5
Multinomial	63.25
Random Forest	87.25
Tuned Random Forest	88.5

Modelling Summary:
In conclusion, the machine learning models, including Logistic Regression (One-vs-Rest), Logistic Regression (One-vs-One), Multinomial Classifier and Random Forest Classifier, were trained and evaluated to classify mobile phones into different price ranges based on various features.
Here are the key findings:

- The Random Forest Classifier achieved the highest accuracy of 88.5% on the test data, followed by the Logistic Regression (One-vs-Rest) with an accuracy of 79.5%, and the Logistic Regression (One-vs-One) with an accuracy of 73.5%.
- All models showed good precision, recall, and F1-score across different price range classes, indicating their ability to effectively classify mobile phones into the correct price categories.
- The Random Forest Classifier, with its ensemble learning approach, demonstrated robust performance and generalisation capability, making it the recommended choice for this classification task.

DEPLOYMENT

Deployment refers to the process of making a software application available for use by its intended users. In the context of software development, deployment involves taking the code or application that has been developed and making it accessible and operational in a production environment where users can interact with it.

There are several key uses and benefits of deployment:

1. Deployment allows software developers to make their applications available to users, whether they are end-users, clients, or other stakeholders.
2. Deployment provides an opportunity to test the application in a real-world environment and validate its functionality, performance, and usability before it is made widely available.

3. Deployment enables end users to access and use the software for its intended purpose, whether it's a web application, mobile app, desktop software, or other types of applications.

Overall, deployment is a critical step in the software development lifecycle that transforms code and software applications into tangible products or services that can be used by individuals, businesses, and organisations to achieve their goals and objectives.

Conclusion and Recommendation

Based on the results, the Random Forest Classifier stands out as the most suitable model for classifying mobile phone price ranges. Therefore, it is recommended to deploy the Random Forest Classifier model for real-world applications, such as mobile e-commerce platforms, where accurate price range classification can assist consumers in making informed purchasing decisions.

Additionally, further optimization and fine-tuning of the Random Forest model could potentially improve its performance even further. This could involve experimenting with different hyperparameters, feature engineering techniques, and data preprocessing methods to enhance the model's accuracy and generalisation ability.

Overall, the developed machine learning model presents a valuable tool for both consumers and businesses in the mobile phone industry, enabling better decision-making and enhancing user experience in the marketplace.