A Machine Learning Regression Analysis

Predicting Smart Phone Prices in India

MACHINE LEARNING AND PATTERN RECOGNITION
Module Code: B9DA109
Professor Name: Mr. Satya Prakash

SUBMITTED BY:

Elisha Johnson Kyanchat
Student ID: 20002405

Peter Chukwuka Ibeabuchi
Student ID: 20007349

Prashanth Periannan
Student ID: 20001940

November 2023.

INTRODUCTION

In today's world, smart phones have become an extension of ourselves. It is an integral part of our everyday activities, from accessing information, to entertainment, and as much as payments and even transportation. Its relevance cannot be overly stated.

For several people, various smart phone specifications/features are more important than others, while some are concerned about camera resolution, others are concerned about battery capacity, RAM size and so on. Which leads to an important question, to what extent does each preferred feature affect the prices of various phone models? Hence understanding the factors that affect these prices is of paramount importance, as it useful for both customers and businesses. This report therefore focuses on accurately predicting the prices of various phone models utilizing machine learning models.

Machine learning, a subset of artificial intelligence, empowers computers to learn from data without explicit programming. By analyzing a vast dataset of smart phone specifications and corresponding prices, machine learning algorithms can identify patterns and relationships that influence the prices of various phone models. This provides customers with more insight as they make the decision on what phone works best for their budget, given the specifications.

This report focuses on accurately predicting the prices of various phone models utilizing machine learning models. The objective is clear: to build a robust and accurate model that estimates the prices of various phone models based on other specifications and features.

METHODOLOGY

In this analysis, we begin by detailing the data preparation process, including data cleaning, and transformation. Next, we carry out a brief descriptive analysis, describing the data and getting important information to help us understand the data better. We also introduce the selection of machine learning algorithms, outlining their strengths and limitations.

The report then presents the evaluation of the developed machine learning model, assessing its performance against various metrics, such as R2_score, Mean Square Error, and Mean Absolute Error. We analyze the factors that influence the model's performance and discuss potential areas for improvement.

DATA PRE-PROCESSING

In this section we examined the dataset to get better understanding of each column. We also cleaned the data, accessing null and duplicate values, and removing unnecessary columns.

The Dataset

The dataset used for this analysis is taken from Kaggle, and can be downloaded here. The dataset contains both numerical and object variables. Here, our primary focus is predicting the prices of phone models, thus the price is the dependent variable, and the rest of the columns are the independent variables. The dataset consists of 22 columns and 1,359 rows. Here is a summary of the columns in the dataset. **It is important to note that the price in this dataset is given in India Rupees (INR)**

Here are the unique columns in the dataset:

- Brand: Brand Name

- Model: Model of the Phone
- Model: Model of the Phone
- Battery capacity (mAh): Battery capacity in mAh
- Screen size (inches): Screen Size in Inches across opposite corners.
- Touchscreen: Whether the phone is touchscreen supported or not.
- Processor: No. of processor cores.
- RAM (MB): RAM available in phone in MB.
- Internal storage: Internal Storage of phone in GB.
- Rear camera: Resolution of rear camera in MP (0 if unavailable).
- Front camera: Resolution of front camera in MP (0 if unavailable).
- Operation system: OS used in phone.
- Wi-Fi: Whether phone has Wi-Fi functionality.
- Bluetooth: Whether phone has Bluetooth functionality.
- Number of SIMs: Number of SIM card slots in phone.
- 3G: Whether phone has 3G network functionality.
- 4G/LTE: Whether phone has 4G/LTE network functionality.
- Price: Price of the phone in INR

	Name	Brand	Model	Battery capacity	Screen size (inches)	Touchscreen	Resolution x	Resolution y	Processor	RAM (MB)	Internal storage	Rear camera	Front camera	Operating system	Wi-Fi	Bluetooth	GPS	Number of SIM	3G	4G/ LTE	Price
0	OnePlus 7T F	OnePlus	7T Pro McLa	4085	6.67	Yes	1440	3120	8	12000	256	48	16	Android	Yes	Yes	Yes	2	Yes	Yes	58998
1	Realme X2 Pi	Realme	X2 Pro	4000	6.5	Yes	1080	2400	8	6000	64	64	16	Android	Yes	Yes	Yes	2	Yes	Yes	27999
2	iPhone 11 Pr	Apple	iPhone 11 Pr	3969	6.5	Yes	1242	2688	6	4000	64	12	12	iOS	Yes	Yes	Yes	2	Yes	Yes	106900
3	iPhone 11	Apple	iPhone 11	3110	6.1	Yes	828	1792	6	4000	64	12	12	iOS	Yes	Yes	Yes	2	Yes	Yes	62900
4	LG G8X ThinQ	LG	G8X ThinQ	4000	6.4	Yes	1080	2340	8	6000	128	12	32	Android	Yes	Yes	Yes	1	No	No	49990
5	OnePlus 7T	OnePlus	7T	3800	6.55	Yes	1080	2400	8	8000	128	48	16	Android	Yes	Yes	No	2	Yes	Yes	34930
6	OnePlus 7T F	OnePlus	7T Pro	4085	6.67	Yes	1440	3120	8	8000	256	48	16	Android	Yes	Yes	Yes	2	Yes	Yes	52990
7	Samsung Ga	Samsung	Galaxy Note	4300	6.8	Yes	1440	3040	8	12000	256	12	10	Android	Yes	Yes	Yes	2	Yes	Yes	79699
8	Asus ROG Ph	Asus	ROG Phone 2	6000	6.59	Yes	1080	2340	8	8000	128	48	24	Android	Yes	Yes	Yes	1	Yes	Yes	37999
9	Xiaomi Redm	Xiaomi	Redmi K20 Pi	4000	6.39	Yes	1080	2340	8	6000	128	48	20	Android	Yes	Yes	Yes	2	No	No	23190
10	Oppo K3	Oppo	K3	3765	6.5	Yes	1080	2340	8	6000	64	16	16	Android	Yes	Yes	Yes	2	Yes	Yes	23990
11	Realme X	Realme	X	3765	6.53	Yes	1080	2340	8	4000	128	48	16	Android	Yes	Yes	Yes	2	Yes	Yes	14999
12	Xiaomi Redm	Xiaomi	Redmi K20	4000	6.39	Yes	1080	2340	8	6000	64	48	20	Android	Yes	Yes	Yes	2	Yes	Yes	19282
13	OnePlus 7 Pr	OnePlus	7 Pro	4000	6.67	Yes	1440	3120	8	6000	128	48	16	Android	Yes	Yes	Yes	2	Yes	Yes	39995
14	Oppo Reno 1	Oppo	Reno 10x Zoo	4065	6.6	Yes	1080	2340	8	6000	128	48	16	Android	Yes	Yes	Yes	2	Yes	Yes	36990
15	Realme 3 Pro	Realme	3 Pro	4045	6.3	Yes	1080	2340	8	4000	64	16	25	Android	Yes	Yes	Yes	2	Yes	Yes	13999
16	Huawei P30	Huawei	P30 Pro	4200	6.47	Yes	1080	2340	8	8000	256	40	32	Android	Yes	Yes	No	2	Yes	Yes	54280
17	Redmi Note 7	Xiaomi	Redmi Note 7	4000	6.3	Yes	1080	2340	8	4000	64	48	13	Android	Yes	Yes	Yes	2	Yes	Yes	9799
18	Huawei Mat	Huawei	Mate 20 Pro	4200	6.39	Yes	1440	3120	8	6000	128	40	24	Android	Yes	Yes	Yes	2	Yes	Yes	63990
19	LG V40 ThinQ	LG	V40 ThinQ	3300	6.4	Yes	1440	3120	8	6000	128	12	8	Android	Yes	Yes	Yes	2	Yes	Yes	29999
20	OnePlus 6T	OnePlus	6T	3700	6.41	Yes	1080	2340	8	6000	128	16	16	Android	Yes	Yes	Yes	2	Yes	Yes	31999
21	Apple iPhone	Apple	iPhone XR	2942	6.1	Yes	828	1792	6	3000	64	12	7	iOS	Yes	Yes	Yes	2	Yes	Yes	45499
22	Apple iPhone	Apple	iPhone XS M	2658	6.5	Yes	1242	2688	6	4000	64	12	7	iOS	Yes	Yes	Yes	2	Yes	Yes	69999
23	Apple iPhone	Apple	iPhone XS	2658	5.8	Yes	1125	2436	6	4000	64	12	7	iOS	Yes	Yes	Yes	2	Yes	Yes	59999
24	Google Pixel	Google	Pixel 3 XL	3430	6.3	Yes	1440	2960	8	4000	64	12.2	8	Android	Yes	Yes	Yes	1	Yes	Yes	47990
25	Google Pixel	Google	Pixel 3	2915	5.5	Yes	1080	2160	8	4000	64	12.2	8	Android	Yes	Yes	Yes	1	Yes	Yes	37999
26	Asus ROG Ph	Asus	ROG Phone	4000	6	Yes	1080	2160	8	8000	128	12	8	Android	Yes	Yes	Yes	2	Yes	Yes	69999
27	Samsung Ga	Samsung	Galaxy Note	4000	6.4	Yes	1440	2960	8	6000	128	12	8	Android	Yes	Yes	Yes	2	Yes	Yes	56999
28	LG G7+ ThinQ	LG	G7+ ThinQ	3000	6.1	Yes	1440	3120	8	6000	128	16	8	Android	Yes	Yes	Yes	2	Yes	Yes	37999
29	Asus ZenFon	Asus	ZenFone Ma	5000	5.99	Yes	1080	2160	8	3000	32	13	8	Android	Yes	Yes	Yes	2	Yes	Yes	9990
30	Huawei P20	Huawei	P20 Pro	4000	6.1	Yes	1080	2240	8	4000	64	40	24	Android	Yes	Yes	Yes	2	No	Yes	49990

Figure 1: By Author- The First Rows of the dataset

Data Processing

- Loading the data in a pandas' data frame: To load the data, we use the panda's function to read csv file, and store the created data frame in a variable.

- Dropping unnecessary columns: The unnamed column has no meaning to it and is therefore not important to our analysis. Also, the Name column is redundant as the information it carries is already contained in the 'Brand' and 'Model' columns, thus, we do not need it. For this analysis we will be dropping both columns.

- Exploring the data types of each column. For better understanding of our dataset, it is important to get all necessary information of the data. After observation, this dataset has 1,359 rows and 20

columns, with no missing values. It contains three data types: four float columns, seven integer columns, and nine object columns.

EXPLORATORY DATA ANALYSIS(EDA)

For a better understanding of the data, it is necessary to understand the description of the dataset. The descriptive analysis gives us a better understanding of the dataset. Here, we discuss certain findings and observations that are significant to our analysis.

Univariate Data Analysis

- **Analyzing popular brands in the dataset.**
  It is important for both customers and businesses to know what brands are most popularly demanded among users. This is helpful for making vital production or purchase decisions.

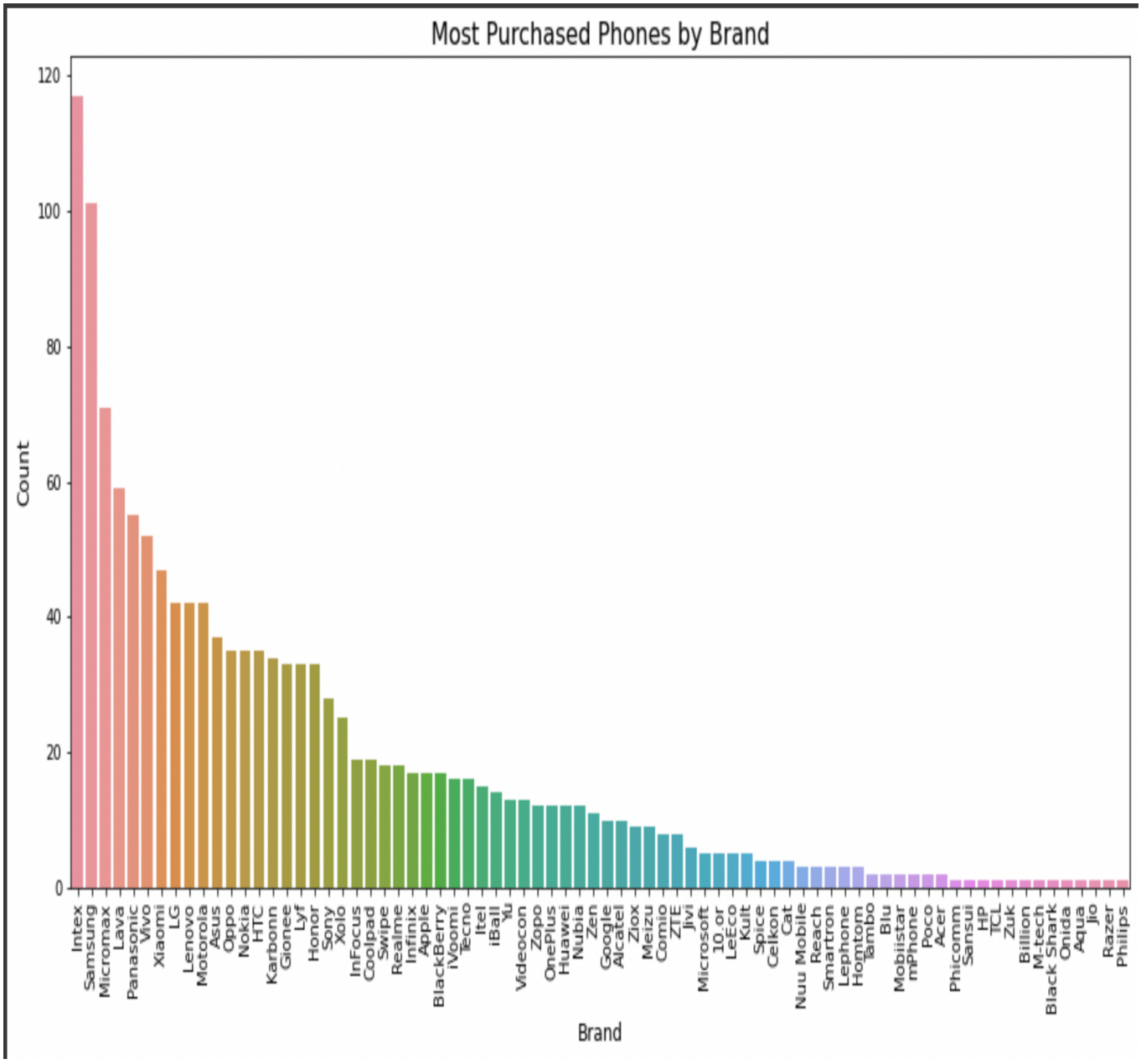


Figure 2: By Author- The Most Purchased Phone Brands

From the chart above we can see that the top 5 most purchased brands in the dataset are Intex, Samsung, Micromax, Lava and Panasonic phone, and the least purchased phones are Onida, Aqua, Jio, Razer and Philips.

Analyzing the Most Patronized Operating System:
We've seen what devices most people subscribe to; however, we need to access what operating systems are also popular in the dataset.

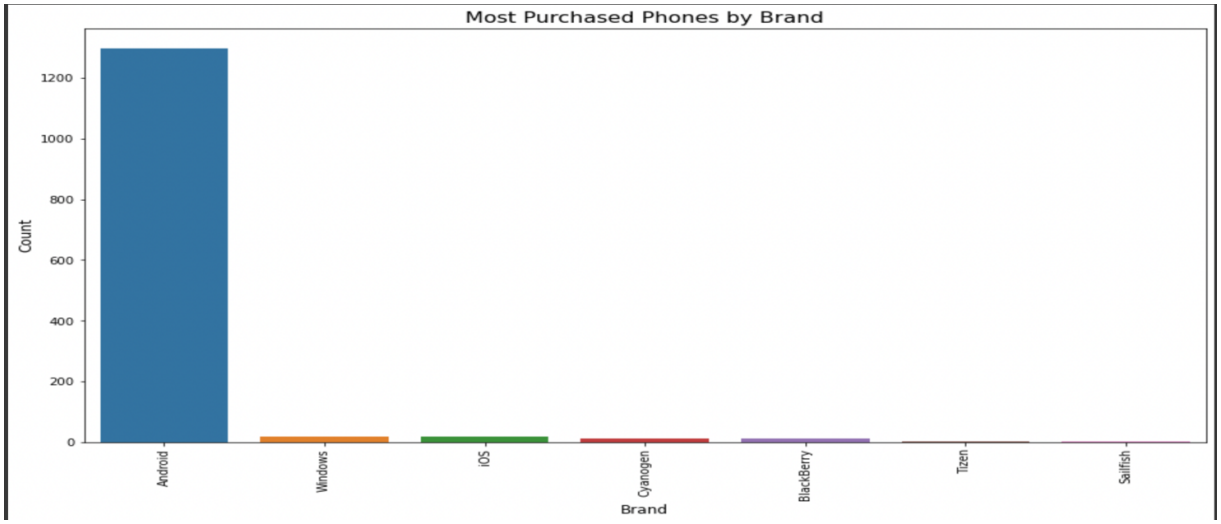


Figure 3: By Author- The Most Popular Operating Systems

Observation: From this, we see that a very good percentage of phones use the Android operating system. Interestingly, the topmost popular phones in the dataset use the Android OS.

- **Categorical Variable Analysis:**

Here, we extract all columns with that are categorical variables (i.e., columns with yes and no categories). We are analyzing the percentage and count of these columns and observing their distribution throughout the dataset.

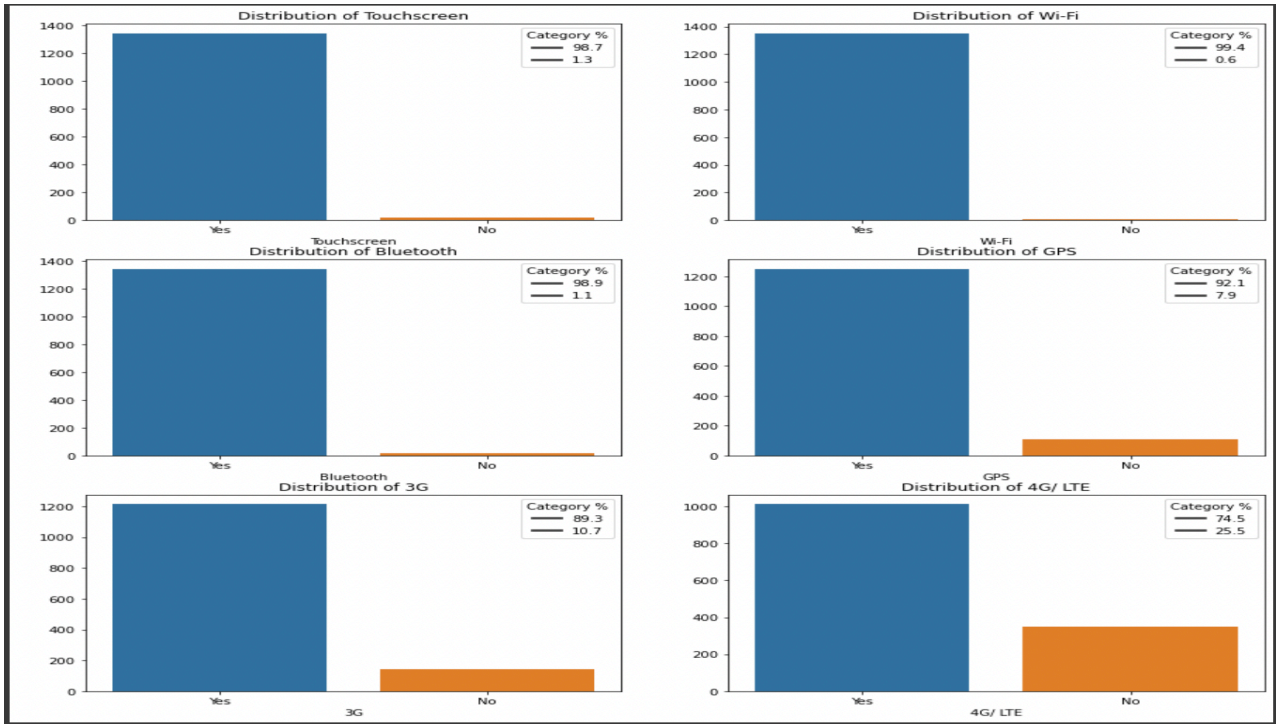
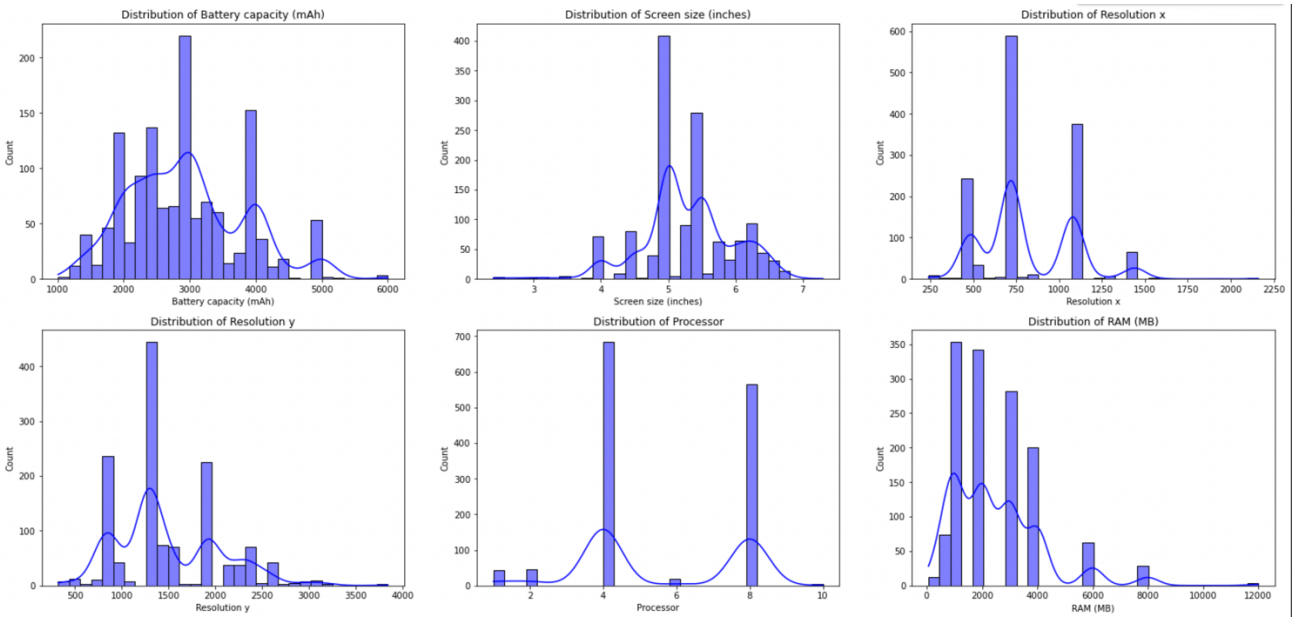


Figure 4: By Author- Distribution of Categorical Variables

From the chart above, we see the following:

- About 99% of the phones have a Touchscreen, Wi-Fi, and Bluetooth
- About 92% of the phones have GPS
- About 90% of the phones have 3G, but only 75% of the phones have 4G/LTE.

- **Numerical Variables Analysis:** In this section we will get the general statistics of the numerical columns. It is important that we visualize the distribution of these features and make critical observations to get more understanding about the data.



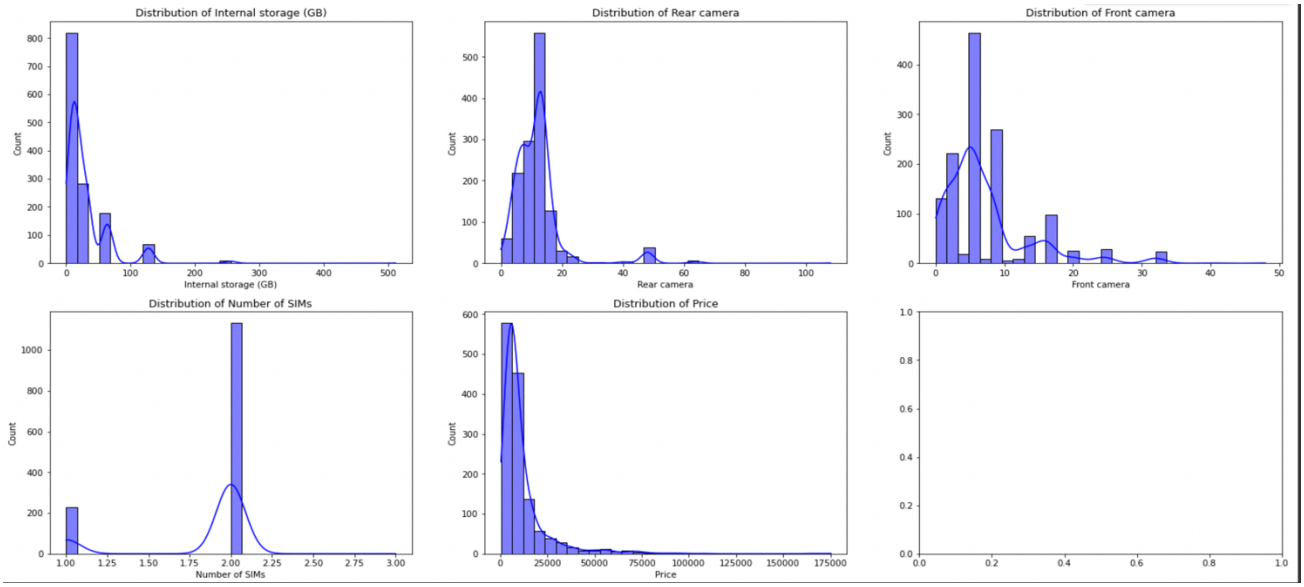


Figure 5: By Author- Distribution of Numerical Variables

From this analysis we can tell the following:

- **Battery Capacity**: The average battery capacity is approximately 2938 mAh, with a minimum of 1010 mAh and a maximum of 6000 mAh. The standard deviation is around 873, indicating some variation in battery sizes.
- **Screen Size:** The average screen size is about 5.29 inches, with the smallest screen being 2.4 inches and the largest being 7.3 inches.
- **Resolution (x and y):** The average resolution for both the x and y dimensions is 811 x 1490 pixels, with some variation. The maximum resolution for both dimensions is considerably higher, indicating some devices with high-resolution screens.
- **Processor:** The average number of processor cores is approximately 5.55, with a minimum of 1 core and a maximum of 10 cores. This suggests a range of processing power among the devices.
- **RAM:** The average RAM is around 2488 MB (or approximately 2.49 GB). The range of RAM sizes varies from 64 MB to 12,000 MB (or 12 GB).
- **Internal Storage:** The average internal storage is approximately 30.65 GB, with a minimum of 0.064 GB and a maximum of 512 GB.
- **Camera (Rear and Front):** The rear camera has an average of 12.07 megapixels, with a minimum of 0 megapixels (possibly indicating missing data) and a maximum of 108 megapixels. The front camera has an average of 7.04 megapixels, with a minimum of 0 megapixels and a maximum of 48 megapixels.
- **Number of SIMs:** On average, devices have approximately 1.83 SIM slots, suggesting that most devices support dual SIM cards. This is also very evident in the "Number of sims" distribution.
- **Price:** The average price for phone in the dataset is approximately 11,465.83 rupee. Prices vary widely, with the minimum at 494 rupees and the maximum at 174,990 rupees. From the phone distribution we can see that most phones cost below 25000 rupees and very few phones cost above 50,000. To see this clearly, we create a range.

Bivariate Analysis:
This is utilized to analyze and explore the data between two datasets or two different variables.

- **Most expensive phone models and operating system:** Here, we compared phone models with the price and operating system with the price to check which model is most expensive and most used operating system.

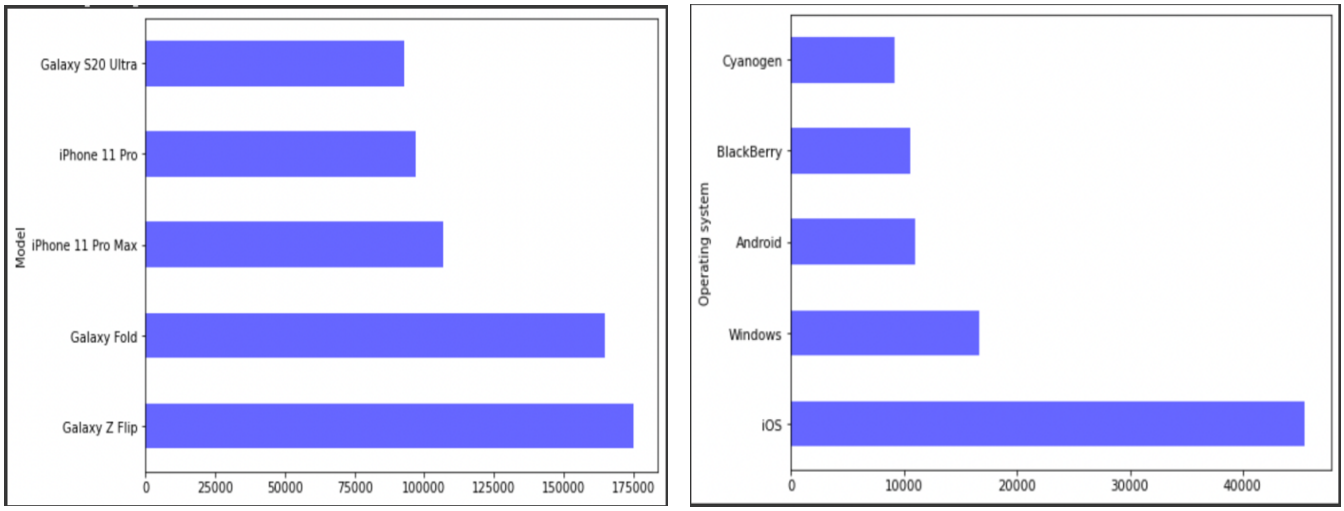


Figure 6: By Author: Comparing Model and OS Prices

From the chart above, we make the following observations.

- On an average, the topmost pricey phone models are Galaxy Z flip, Galaxy fold and iPhone 11 pro max, and the cheapest phone are the U12 and Mate 30 pro.
- iOS phones seem to be the costliest phones in the dataset, followed by windows.

PRE-PROCESSING

Here, before we pass the data to the model, we are trying to make sure that the data is clean (removing missing values, duplicates, etc.) and numerically encoded (one hot encoding)

LABEL-ENCODING

Label-Encoding is the process of changing the categorical variable into numerical format as we require numerical format to run models in machine learning.

To make the dataset compatible with machine learning model, we hot encoded all object variables to make them numerical. Here is what our new table looks like now.

	Brand	Model	Battery capacity (mAh)	Screen size (inches)	Touchscreen	Resolution x	Resolution y	Processor	RAM (MB)	Internal storage (GB)	Rear camera	Front camera	Operating system	Wi-Fi	Bluetooth	GPS	Number of SIMs	3G	4G/LTE	Price
0	44	49	4085	6.67	1	1440	3120	8	12000	256.0	48.0	16.0	0	1	1	1	2	1	1	58998
1	53	1142	4000	6.50	1	1080	2400	8	6000	64.0	64.0	16.0	0	1	1	1	2	1	1	27999
2	3	1288	3969	6.50	1	1242	2688	6	4000	64.0	12.0	12.0	6	1	1	1	2	1	1	106900
3	3	1286	3110	6.10	1	828	1792	6	4000	64.0	12.0	12.0	6	1	1	1	2	1	1	62900
4	29	522	4000	6.40	1	1080	2340	8	6000	128.0	12.0	32.0	0	1	1	1	1	0	0	49990

Figure 6: By Author - The First Rows of the Dataset

FEATURE SELECTION

To optimize the model's performance, it is important to identify those features that have better correlation with the target variables (Price). We achieve this by picking those models that have at least 10% correlation with price.

Here is a visual representation of the correlations:

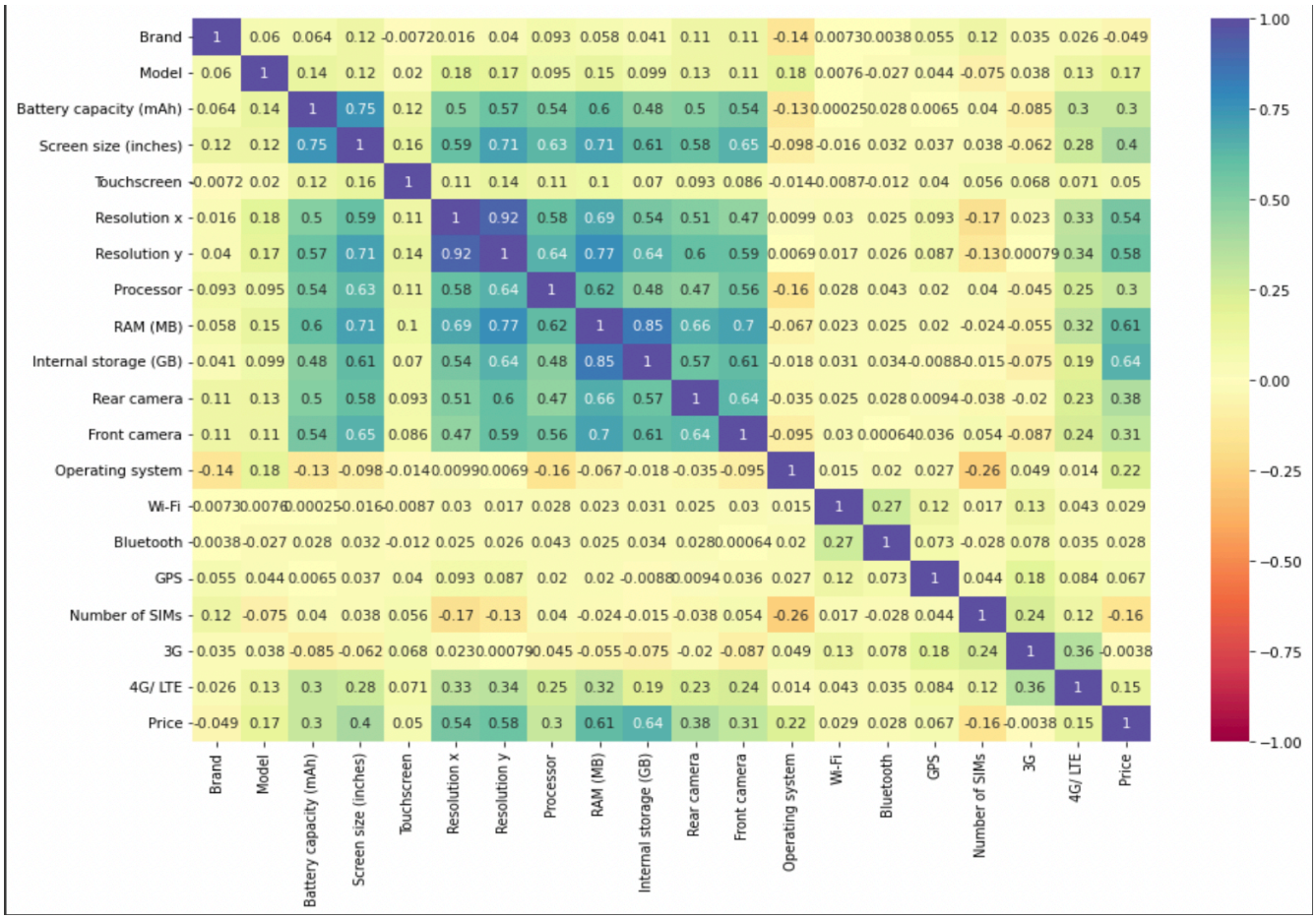


Figure 7: By Author- Correlation Map of all Columns in the Dataset

From this, it is evident that there are quite a few correlations between most smartphone features. Let's highlight the most significant.

- The strongest positive correlation is between screen size and Resolution y (r =0.92), meaning that there is a very strong linear relationship between the two features.
- Battery capacity is positively correlated with screen size, resolution, processor speed, RAM, internal storage, and rear camera resolution. This means that phones with larger batteries tend to have the best of these features. However, we that the battery capacity has a very low correlation with Number of sims, showing that this feature does not have much impact on the capacity of a phone's battery.
- The number of sim have very low correlation to all the features.
- Internal storage has a height positive relationship with Screen Size, Resolution y, RAM, Front Camera, and Price.
- The RAM of phones has a high positive relationship all features safe for "Number of sims"
- Price has a high positive relationship with the phones Resolution (both X and Y), the RAM, and internal storage.

- Here we see a positive relationship between WIFI and Bluetooth, and between 4G/LTE and 3G.
- 4G/LTE also shows a good positive correlation with Resolution (X and Y), as well as RAM.

MODEL SELECTION

Selecting the appropriate model is important for building an accurate predictive model. The choice of the model in largely dependent on the type of data being analyzed. Given our dataset in this project, we are using regression models, and to achieve better accuracy, we have chosen three models. Each of these models are unique and bring a specific strength for our analysis.

- **Linear Regression:** Linear Regression is a foundational model widely used for predicting numeric values. Here we consider only one independent variable to find relationship between dependent variable i.e., Price, which remains unchanged and to identify the linear path that finest matches the data and figure out the ideal intersection and coefficient values to reduce error.

  Y=mX+C

  Where, 'Y' is our target value which is to be predicted, 'X' is independent variable, 'm' is the slope and 'C' is the intercept.

- **XGBoost Regressor:** It stands for Extreme Gradient Boosting. Here, the main function of XGBoost is to predict continuous numeric value and dominates structured or tabular datasets on regression predictive modeling problems. This model offers high predictive accuracy and is capable of complex interaction between features.

- **RandomForest Regressor**: This model is an ensemble of decision trees. Here, a group of decision trees is constructed, with each tree anticipating a numerical result on each of them. The result is derived by combining the predictions from each of the individual trees. It is best suited for robust overfitting and is also capable of interacting with complex relationships.

MODEL SELECTION STRATEGY:

To determine the most suitable model for our smartphone price prediction, we will undertake a comprehensive evaluation process. This includes assessing each model's performance based on key metrics such as R-Squared(R2), Mean Absolute Error (MAE), and Mean Squared Error (MSE).

- **R2Score:** It's a statistical metric that represents the goodness of fit of a regression model.
- **Mean Squared Error (MSE):** It's calculated by taking mean or average of squared difference from data and the function.
- **Mean Absolute Error (MAE):** the amount by which an observation's true value differs from its prediction.

MODEL OPTIMIZATION:

To enhance the accuracy of our predictive model, we employed several optimization techniques. These techniques involved standardizing the data using the MinMaxScaler() and StandardScaler() normalization methods. Additionally, we trained our model using selected variables (features) that demonstrated a stronger correlation with prices. This process is well detailed in the feature selection section.

MinMaxScaler(): This function transforms the data by scaling each feature to a specific range, typically between 0 and 1. This standardization technique helps to ensure that all features are on a comparable scale, preventing features with larger magnitudes from dominating the model's training process.

StandardScaler(): This function normalizes the data by subtracting the mean of each feature from each data point and then dividing by the standard deviation. This normalization technique centres the data around a mean of 0 and scales it to a standard deviation of 1, ensuring that all features have equal influence on the model's training process.

Feature Selection: Feature selection involves identifying and selecting the most relevant features from the dataset. This process helps to reduce the dimensionality of the data, which can improve model performance by eliminating irrelevant or redundant features. In our case, we selected features that exhibited a stronger correlation with prices, ensuring that the model focuses on the most influential factors.

MODEL COMPARISON AND EVALUATION

We have now carried out several test to get the best models, both training and testing the plain dataset, to training and testing standardized/normalized dataset, as well as training and testing only selected features in the dataset. So many models have performed fairly, some more than other. Here is a table showing the combined list of each model's performance.

Model	R2_Score(%)	MAE	MSE
LR_metrics(Plain)	64.0	5188.76	95015467.31
XGB_metrics(Plain)	63.0	4064.00	97234363.57
Rf_metrics(Plain)	69.0	3778.26	81744478.39
LR_metrics(FS)	64.0	5176.67	96171676.06
XGB_metrics(FS)	67.0	4212.18	88041595.67
RF_metrics(FS)	67.0	3925.13	85684340.13
LR_metrics(SD)	64.0	5188.76	95015467.31
XGB_metrics(SD)	63.0	4064.00	97234363.57
RF_metrics(SD)	66.0	3834.36	88494962.23
LR_metrics(SF)	64.0	5176.67	96171676.06
XGB_metrics(SF)	67.0	4212.18	88041595.67
RF_metrics(SF)	68.0	3888.91	83564080.14
LR_metrics(MM)	64.0	5188.76	95015467.31
XGB_metrics(MM)	63.0	4064.00	97234363.57
RF_metrics(MM)	67.0	3929.93	85805653.58
LR_metrics(MF)	64.0	5176.67	96171676.06
XGB_metrics(MF)	67.0	4212.18	88041595.67

Figure 8: By Author- Metrics Table

Considering the overall performance across various preprocessing techniques, the RF_metrics(MF) appears to be the most robust and accurate, achieving the highest R2 score and demonstrating the lowest MAE and MSE. However, these scores can be improved on, and our models can be optimized to produce even better results.

CONCULSION

In this study, we investigated the effectiveness of machine learning in predicting smartphone prices. We evaluated three distinct models: linear regression, XGBoost, and random forest. Through several testing and optimization techniques, we demonstrated that random forest outperformed the other models, achieving an accuracy of 69% on the testing dataset. However, these results can further be improved on.

We recommend exploring the use of more complex machine learning algorithms and incorporating additional data sources, such as user reviews and market trends to get better analysis. Also, optimization methods like tunning the parameters of each model can greatly improve the performances.

REFERENCES:

Agarwal, A. (2023, October 12). Machine Learning - Label Encoding of Datasets in Python. Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/

Brownlee, J. (2020, October 10). Feature Selection Techniques in Machine Learning. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/

IBM. (2023, August 10). What is Machine Learning? Retrieved from: https://www.ibm.com/topics/machine-learning